Radiation Sensors

Physical Foundations

JOHAN BLEEKER
SRON Netherlands Institute for Space Research
Astronomical Institute Utrecht

Contents

Chapter 1

Transducers: general aspects

1.1 Information handling systems

Information is processed and manipulated in *information handling* systems, which should eventually result in:
-accessibility for human sensing
-control of the information status
The information handling can generally be subdivided in three consecutive stages, i.e. identification, processing/modification and presentation. Each of these stages can be separately characterized as follows:
- *The identification stage* transforms the incoming information from one particular physical form into another with the aid of a so-called transducer. Since in this case it regards the input of the information handling system it is designated as input(i/p) transducer. In the large majority of cases the output of this stage is in the form of electrical energy that reflects the behavior of the original information carrier. Examples comprise a thermocouple, a pH-measurement cell, a TV-camera.
- *The modification stage* does not alter the physical form of the information carrier but processes and adapts the information to provide a suitable interface with the output stage for presentation. In practice the modification stage comprises a set of electronic (integrated) circuitry (IC) on chips and printed circuit boards. Prominent example: an Analogue to Digital Converter (ADC).
- *The presentation stage* transforms the information processed by the modification stage into a physical form that is accessible for human sensing, i.e. information display. Yet again this involves a transducer and since it now concerns the output of the information handling system it is designated as output(o/p) transducer. However in this case the application is much more widespread than only display: the information output may be used in control systems (i.e. implementation) or, in case the information is not immediately needed for interpretation, it can be stored for later usage (storage). In addition, if the information should be made available elsewhere, there may be a need for transmission (data link). Figure (1.1) shows at the top the subdivision in three stages, along with the listing of alternative forms of information energy. Also the diagram displays representations in hardware elements (middle picture) and functional units at the bottom. Regarding the hardware elements the i/p transducer embodies in practice

a sensor or a sensing element that is capable of detecting a physical or chemical phenomenon, e.g. a light spectral intensity, a mechanical displacement, a temperature, a magnetic field or an acidity, and of converting such a signal in most cases in an electrical entity. Turning now to the o/p transducer and taking transmission as an example, it may constitute an antenna array or a glass fibre cable.

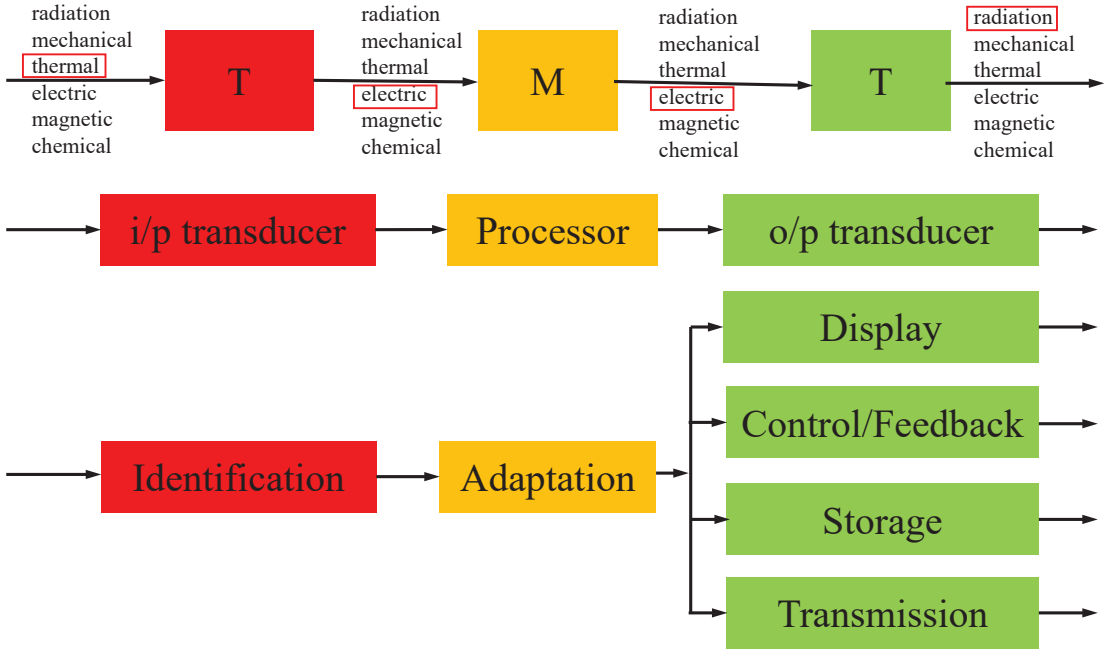


Figure 1.1: *Instrument systems: different forms of signal carrying energies, subdivision in hardware elements and functional units.*

1.2 Transducers

The form of energy in which the information carrier manifests itself at the input of the i/p transducer can take one of the following shapes:
- *Radiation energy*: EM-radiation from radio waves to γ-ray photons.
- *Gravitational energy*: Interaction between masses as a consequence of gravitational pull, i.e. tidal forces.
- *Mechanical energy*: displacements, pressure, acceleration, torque, flows etc.
- *Thermal energy*: kinetic energy of atoms and molecules. either free (gasses) or bound (phonons in solids).
- *Electromagnetic energy*: $\vec{E}$ and $\vec{B}$ fields, voltage, current.
- *Molecular energy*: energy associated with molecular binding forces like the van der Waals force.
- *Atomic energy*: energy associated with electron binding forces at the atomic nucleus.
- *Nuclear energy*: energy associated with binding forces between nucleons.
- *Mass energy*: $E = \gamma mc^2, \gamma \geq 1$.

In principle all of these forms of energy could act as an information carrier, however

in \ out	radiative	mechanical	thermal	electrical	magnetic	chemical
radiative	photo-luminiscence	radiation pressure	radiation heating	photo-conduction voltaic effect	photo-magneto-electric effect	photo-chemical effect
mechanical	photo-elastic effect	momentum exchange	friction heat	pièzo-electric effect	magneto-striction	pressure-induced explosive
thermal	incande-scence	thermal expansion	heat conduction	Seebeck effect	Curie-Weiss law	endothermal reaction
electrical	electro-luminiscence	pièzo-electric effect	Peltier effect	pn-junction effect	Ampère's law	electrolysis
magnetic	Faraday effect	magneto-striction	Ettinghausen effect	Hall effect	magnetic induction	
chemical	chemo-luminiscence	explosive reaction	exothermal reaction	Volta effect		chemical reaction

Figure 1.2: *Matrix showing physical/chemical effects for signal and energy conversion for six different domains as elaborated in the text.*

in most cases a certain grouping (or splitting) takes place (grouping of mechanical and gravitational energy, splitting of electrical and magnetic energy, disregarding nuclear and mass energy) that leads to the following six nominal signal domains:

- Radiative signal domain
- Mechanical signal domain
- Thermal signal domain
- Electrical signal domain
- Magnetic signal domain
- Chemical signal domain

Figure (1.2) shows a table that explicates several physical and chemical effects for signal/energy conversion between these six domains. The six signal domains are also highlighted in figure (1.1) at the in- and output terminals of the consecutive stages.
There are several possibilities to categorize the i/p transducer. In practice three categories are mostly encountered:
1. *Information oriented*, i.e. the transducer provides information about a certain subject, e.g. a traffic situation, climate, health, a TV-broadcast. Consequently the relevant

transducers are categorized according to discipline, e.g. feedback and control, meteorological, biomedical etc.
2. *Measurand oriented*, i.e the transducer is labeled in line with the quantity that needs to be measured. Example: the measurand is a movement that needs to be assessed for cars, bicycles and pedestrians at a particular traffic location. Such a movement can be physically measured in several different ways: optically, mechanically, electrically or thermally (all potential information carriers).
3. *Parameter oriented*, i.e. the transducer is designed to measure a physical or chemical parameter, embodied by the information carrier itself, for example an amount of light, a change in pressure or a difference in temperature.
For many i/p transducers hold that their operation is based on a single physical effect like e.g. the Seebeck effect (employed in a thermocouple) or the pièzo-electric effect. Transducers that employ multiple physical effects for their operation are designated tandem-transducers. Example: measuring the speed of a gas flow (measurand) by employing two physical effects (i) a temperature drop incurred through cooling by the gas flow and (ii) employing the Seebeck effect for conversion of the temperature drop into an electrical signal.
Apart from the subdivision in single-element and tandem-transducers, one distinguishes between:

- *Self generating* transducers.

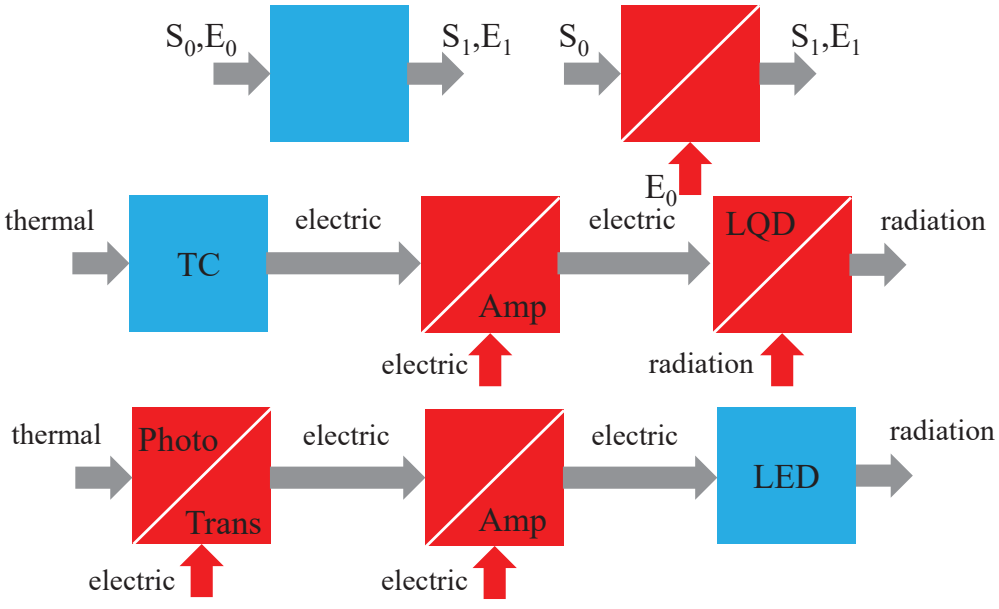


Figure 1.3: *Symbols for self-generating and modulation transducers (picture at the top). In the middle and at the bottom: two alternative transducer configurations for a digital thermometer, employing different input transducers (thermocouple(TC) and phototransistor) and different output transducers (Liquid Crystal Display(LQD) and Light Emitting Diodes(LED)).*

In this case, apart from the input signal, no external energy source is required for the operation of the transducer. Examples constitute thermocouples and solar cells.

- *Modulated* transducers.
  In this case the transducer is energized by an external power source and, on top of this, is modulated by the physical or chemical entity that needs to be measured. Examples include photoconductors, pressure cells equipped with a pièzo resistance element.

Both type of sensors are often represented by a different symbol as shown at the top in figure (1.3). Also indicated are two alternative solutions for a digital thermometer design, employing different transducer elements, one design invokes a self-generating i/p transducer (i.e. a thermocouple), the other design invokes a self-generating o/p transducer (i.e. a LED display).

1.3 Sensors: performance parameters

In characterizing sensor performance the following parameters can be considered:

- *Accuracy* : The ratio of error to output.
- *Creep* : A slow and continuous change in output.
- *Directivity* : The angle dependence of the sensitivity.
- *Drift* : A time dependent change of the output signal.
- *Hysteresis* : The difference in output at a certain value of the measurand when this value is approached first with an increasing and next with a decreasing measurand.
- *Linearity* : The closeness of a calibration curve to a specified straight line.
- *Offset* : The output signal of the sensor when the measurand is zero.
- *Output impedance* : The impedance across the output terminals of the sensor.
- *Range* : The values between the upper and the lower limits for which the sensor is intended.
- *Reliability* : A measure of the probability that the sensor will operate for a specified length of time.
- *Repeatability* : The ability of a sensor to generate the same output when the same measurand value is applied.
- *Resolution* : The smallest step-like change of the measurand that can be detected.
- *Resonant frequency* : The frequency at which the sensor shows maximum output.
- *Response time* : The time required for the sensor output to reach its final value.
- *Sensitivity* : The ratio between the change in sensor output and the value of the measurand.
- *Transfer function* : The relation between the sensor output and the measurand.
- *Transient response* : The response of the sensor to a step-like change in the measurand.

Chapter 2

Sensor interface: equivalent circuits

2.1 Equivalent circuits and optimal power adaptation

In practical situations any influence of the sensor's presence on the measurand should be avoided, i.e. in many cases the aim is miniaturization. This often leads to small signal levels, hence amplification and optimum frequency filtering are mandatory to counteract the influence of external disturbances to maintain/acquire an acceptable signal to noise ratio (S/N).
An important notion in this context is the *maximum* available power that the sensor can deliver to the processing chain. This requires *optimum* power adaptation at the sensor

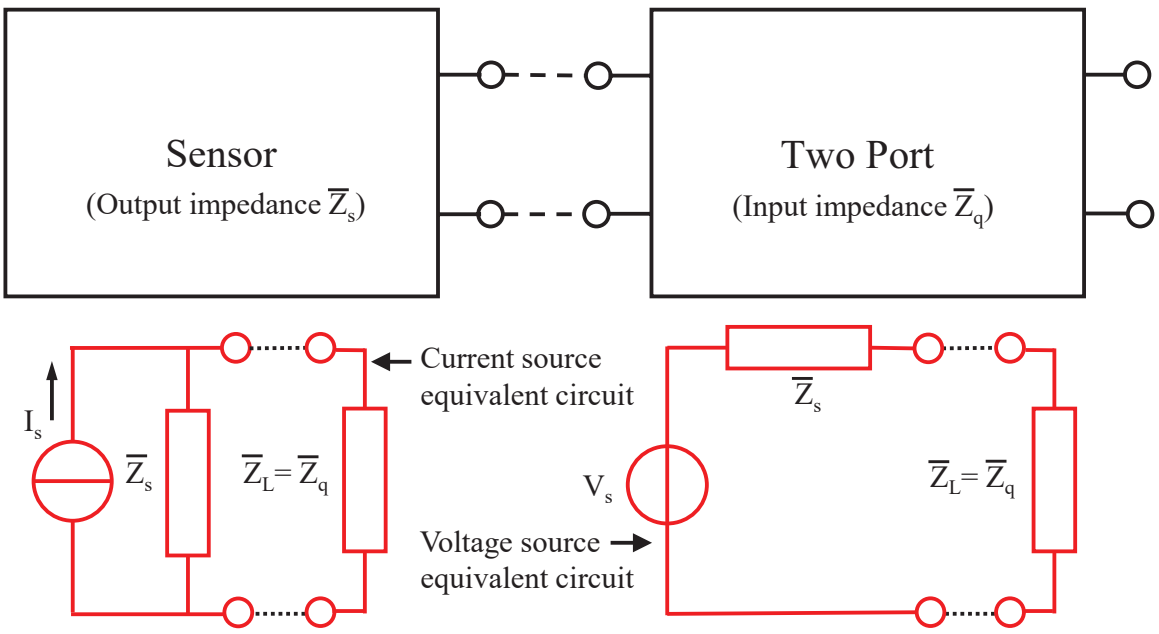


Figure 2.1: *Loading of the sensor output. Left: schematic of an equivalent current source, Right: schematic of an equivalent voltage source.*

output by matching the input impedance $\bar{Z}_q$ of the first stage two port of the processor chain to the output impedance $\bar{Z}_s$ of the sensor. Let's first assume for simplicity that $\bar{Z}_s$ and $\bar{Z}_q$ can be regarded as pure resistive elements, i.e. $\bar{Z}_s = R_s$ and $\bar{Z}_q = R_q$. This situation is depicted in figure 2.1, which shows a current and voltage equivalent circuit of the electric interface. Evaluating the voltage equivalent circuit, the signal power dissipated by the sensor signal in the load resistor $R_L(= R_q)$ can be expressed as:

$$W_s = \frac{V_s R_L}{(R_s + R_L)} \cdot \frac{V_s}{(R_s + R_L)} = \frac{V_s^2 R_L}{(R_s + R_L)^2} \quad \text{optimum power extraction follows from}$$

$$\frac{\partial W_s}{\partial R_L} = 0 \quad \text{yielding} \quad R_L = R_s, \quad \text{hence the available power from the signal source equals:}$$

$$W_s = \frac{V_s^2}{4R_s}, \quad \text{also} \quad W_s = \frac{I_s^2 R_s}{4} \tag{2.1}$$

Let us now consider the more general case of complex impedances $\bar{Z}_s = R_s + jX_s$ and $\bar{Z}_q = R_q + jX_q$, in which X_s and X_q represent the impedance *reactances*. We can then express the power P as a complex entity $P = W + jQ$. The real power W is the average power [Watt] delivered to a load and represents the only useful power: it is the actual power dissipated by the resistive element (Ohmic heating) of the load. The *reactive* power Q is a measure of the energy exchange between the source and the reactance part of the load and does not contribute to any power dissipation. Hence the real power dissipated in the load impedance $\bar{Z}_L$ $(= \bar{Z}_q)$ becomes:

$$W_s = \frac{V_s^2 R_L}{\left(|\bar{Z}_s + \bar{Z}_L|\right)^2} = \frac{V_s^2 R_L}{(R_s + R_L)^2 + (X_s + X_L)^2} \quad \text{optimum power extraction from:}$$

$$\frac{\partial W_s}{\partial R_L} = 0 \quad \text{and} \quad \frac{\partial W_s}{\partial X_L} = 0 \quad \text{yielding} \quad R_L = R_s \quad \text{and} \quad X_s = -X_L \tag{2.2}$$

Apparently optimal power transfer will occur if the load impedance $\bar{Z}_L$ and the source impedance $\bar{Z}_s$ do have the same real part, whereas the imaginary parts should have the same value but opposite signs. Hence, $\bar{Z}_L$ and $\bar{Z}_s$ are called *conjugate* impedances, i.e. $\bar{Z}_L = \bar{Z}_s^*$. Since for the case of optimal adaptation the complex parts of $\bar{Z}_L$ and $\bar{Z}_s$ cancel, expression (2.1) also holds for the maximum available power, now with R_s representing the real part of impedance $\bar{Z}_s$.

2.2 Characterization of noise, the Wiener-Khinchin relation

Noise signals generated by a sensor and its associated electronic processing chain can be described by a stochastic variable $V(t)$ with Gaussian probability density distribution $\mathbf{p}(V(t))$ for each value of t, centered at zero mean with a mean square deviation:

$$\mathbf{E}\{V^2(t)\} = \int_{-\infty}^{+\infty} V^2(t)\mathbf{p}(V(t))dV(t) = \sigma^2(t), \text{the } \textit{variance} \text{ of the Gaussian at } t \tag{2.3}$$

This can be written as $\sigma^2(t) = \overline{V^2}(t)$, a derived quantity is the *standard deviation* $\sqrt{\overline{V^2}(t)}$, which is a measure of the noise level strength and commonly referred to as the rms-noise amplitude. The noise signal $V(t)$ is stationary if the statistical properties of the signal do not vary with time. This implies that the value of the variance $\overline{V^2}(t) = \overline{V^2} = \sigma^2$, independent of time. As long as macroscopic entities like temperature, resistor values and frequency bandwidth do not change, $V(t)$ will remain stationary. Moreover, a stationary stochastic process is considered *mean ergodic* if the momentaneous *ensemble* average over $V(t)$ can be interchanged with its *time* average, i.e. when:

$$\mathbf{E}\{V(t)\} = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{1}{2}T}^{+\frac{1}{2}T} V(t)dt \tag{2.4}$$

In the electronics discipline it is customary to use the term *power* for the variance σ^2 and the term *effective value* for the standard deviation. This might sound misleading since, assuming that $V(t)$ is a noise *voltage*, the dimensions are [Volt^2] (instead of [Watt]) and [Volt] (instead of [Watt$^{\frac{1}{2}}$]) respectively. Nevertheless σ^2 is proportional to the genuine electric power that a noise voltage can deliver into a load. If $V(t)$ is a noise$-$emf with variance σ^2, the average power delivered to a load resistor R equals σ^2/R [Watt].
The *one sided* (voltage) power spectral density $S_V(\nu)$, with dimension [Volt^{2}Hz^{-1}], represents the distribution of the total average power σ^2 over all physical frequencies ($0 < \nu < +\infty$ = one-sided!):

$$\sigma^2 = \int_0^{+\infty} S_V(\nu)d\nu \tag{2.5}$$

Consider feeding a deterministic voltage signal $\bar{V}_x = |\bar{V}_x|e^{2\pi j\nu t}$ into a filter circuit with a complex Modulation Transfer Function (MTF) $\bar{H}(2\pi j\nu) = |\bar{H}(2\pi j\nu)|e^{j\phi}$, the output voltage $\bar{V}_y$ is then given by:

$$\bar{V}_y = \bar{H}(2\pi j\nu)\bar{V}_x = |\bar{H}(2\pi j\nu)|e^{j\phi} \cdot |\bar{V}_x|e^{2\pi j\nu t} = |\bar{H}(2\pi j\nu)||\bar{V}_x|e^{j(2\pi\nu t+\phi)} \tag{2.6}$$

in which $|\bar{H}(2\pi j\nu)|$ represents the *amplitude* amplification and ϕ the *phase* shift introduced by the frequency filter. In the case of a noise signal, its randomly fluctuating phase is irrelevant, the power amplification of the noise signal follows from $\bar{H}(2\pi j\nu) \cdot \bar{H}^*(2\pi j\nu) = |\bar{H}(2\pi j\nu|^2$, in which $\bar{H}^*(2\pi j\nu)$ is the complex conjugate of the MTF function.
If we now consider a noise signal $V_x(t)$ with a power spectral density $S_{V_x}(\nu)$ [Volt^{2}Hz^{-1}] that is transmitted through the above filter, the power spectral density $S_{V_y}(\nu)$ at the output of the filter becomes:

$$S_{V_y}(\nu) = |\bar{H}(2\pi j\nu)|^2 \cdot S_{V_x}(\nu), \text{ total power } \sigma_y^2 = \int_0^{+\infty} |\bar{H}(2\pi j\nu)|^2 S_{V_x}(\nu)d\nu \tag{2.7}$$

Since the filter will effectively have a high frequency cut-off, say at some value ν_c, expression (2.7) will yield a finite value for the output power as it physically should be!

The structure of the noise signal in the time domain can be assessed with the aid of the autocovariance function $C(\tau)$, for a stationary ergodic signal with *zero* mean, i.e. autocovariance equals autocorrelation, in case of an arbitrary (e.g. voltage, charge or current) time dependent variable $z(t)$ defined as:

$$C(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{1}{2}T}^{+\frac{1}{2}T} z(t) \cdot z(t+\tau)\, dt = \mathbf{E}\{z(t) \cdot z(t+\tau)\} \tag{2.8}$$

By definition $C(\tau)$ is an *even* function, i.e. $C(\tau) = C(-\tau)$, if $\tau \to 0$ we get $C(0) = \sigma^2$. Since the correlation will diminish for increasing τ we have $C(\pm\infty) = 0$. Qualitative interpretation tells: the variance gives the amount of fluctuation in $z(t)$, whereas the autocovariance gives a measure to what extent $z(t)$ and $z(t+\tau)$ track each other, i.e. are correlated. Moreover, the expression for $C(\tau)$ implies that the noise signal can be considered as both *mean* and *correlation* ergodic.
The effective correlation time τ_{eff}, a measure for the global length of the correlation function, determines to a large degree the outlook of the noise signal. One can intuitively understand that if the effective correlation time becomes longer, less fast fluctuations will show up in the noise signal, i.e. the frequency bandwidth of the noise signal is decreased. This should become evident from the inspection of the power spectral density $S_z(\nu)$. In fact there is a very precise relation between $C(\tau)$ and $S_z(\nu)$, they form a Fourier pair $C(\tau) \Leftrightarrow S_z(\nu)$. This Fourier transform is known as the Wiener-Khinchin relation. In Fourier analysis the independent variables range from $-\infty \to +\infty$ for symmetry reasons, so we have $-\infty < \nu < +\infty$ and $-\infty < \tau < +\infty$. The power spectral density $S_V(\nu)$ used earlier was one-sided ($0 < \nu < +\infty$) since it covered only positive physical frequencies. We define now the *double sided* power spectral density $S_{d_z}(\nu) = \frac{1}{2}S_z(\nu)$ for $\nu > 0$ and $S_{d_z}(\nu) = \frac{1}{2}S_z(-\nu)$ for $\nu < 0$, both $C(\tau)$ and $S_{d_z}(\nu)$ in the Fourier pair are then even functions.
For the Wiener-Khinchin theorem we can now write:

$$S_{d_z}(\nu) = \int_{-\infty}^{+\infty} C(\tau) \cdot e^{-2\pi j\nu\tau}\, d\tau \tag{2.9}$$

$$C(\tau) = \int_{-\infty}^{+\infty} S_{d_z}(\nu) \cdot e^{2\pi j\nu\tau}\, d\nu \tag{2.10}$$

In case z represents a voltage we have power dimensions [Volt^2] for $C(\tau)$ and [Volt^{2}Hz^{-1}] for $S_{d_z}(\nu)$, in case z represents a current we have power dimensions [Ampere^2] for $C(\tau)$ and [Ampere^{2}Hz^{-1}] for $S_{d_z}(\nu)$
The theorem is particularly useful for analyzing linear time-invariant systems when the inputs and outputs are not square-integrable.

2.3 Noise (degradation) factor, Friis' cascade rule

The degradation of the signal to noise ratio at the output of the sensor due to the noise introduced by the electronic elements of the processing chain can be expressed in terms of a so-called noise factor (or degradation factor) Δ.
First consider the sensor coupled to a single stage processor (e.g. amplifier) with a transfer function $\bar{H}(2\pi j\nu)$, this implies a spectral power gain $G(\nu) = |\bar{H}(2\pi j\nu|^2$. Let's define a signal spectral density at the sensor output $S_s(\nu)$ and a noise spectral density $N_s(\nu)$, at the output of the processing stage we then have $S_o(\nu)$ and $N_o(\nu)$ with $S_o(\nu) = G(\nu) \cdot S_s(\nu)$. The noise factor is now defined as:

$$\Delta(\nu) = \frac{S_s(\nu)/N_s(\nu)}{S_o(\nu)/N_o(\nu)} \tag{2.11}$$

The output spectral noise power can now be expressed as:

$$N_o(\nu) \;=\; G(\nu)N_s(\nu) + [\Delta(\nu) - 1]G(\nu)N_s(\nu) \tag{2.12}$$

The first term in (2.12) refers to the noise contribution of the sensor, the second term represents the spectral noise power of the processing stage.
Let's consider next a processing chain with n-stages, with spectral gain $G(\nu) = \prod_{i=1}^{n} G_i(\nu)$.
The output noise of this multi-stage chain follows from:

$$\begin{aligned}
N_o(\nu) \;&=\; G_1(\nu), G_2(\nu), G_3(\nu).....G_n(\nu)N_s(\nu) \\
&+\; [\Delta_1(\nu) - 1]G_1(\nu), G_2(\nu), G_3(\nu).....G_n(\nu)N_s(\nu) \\
&+\; [\Delta_2(\nu) - 1]G_2(\nu), G_3(\nu).....G_n(\nu)N_s(\nu) \\
&+\; [\Delta_3(\nu) - 1]G_3(\nu).....G_n(\nu)N_s(\nu) \\
&+\; ..... \\
&+\; ..... \\
&+\; ..... \\
&+\; [\Delta_n(\nu) - 1]G_n(\nu)N_s(\nu), \text{ which can be rewritten as:}
\end{aligned}$$

$$N_o(\nu) \;=\; G(\nu)N_s(\nu)\left[1 + \sum_{i=1}^{n} \frac{[\Delta_i(\nu) - 1]}{\prod_{k=0}^{i-1} G_k(\nu)}\right], \text{ with } G_0(\nu) = 1 \tag{2.13}$$

Expression (2.13) for the power spectral density [Watt Hz^{-1}] of the output noise shows a complicated dependence on the frequency characteristics of the input noise and of the electronic processing chain. Therefore let's introduce a number of simplifying assumptions to allow an assessment of the total power [Watt] of the output noise by defining *a priori* a frequency spectrum of the input noise from the sensor $N_s(\nu)$ and frequency transfer functions of the gain factors $G_i(\nu)$. Select a thermal model for the input noise $N_s(\nu)$: the available spectral noise power equals kT_s [Watt Hz^{-1}] and is independent of frequency, so-called white noise. The physics behind thermal noise sources will be

treated in the following chapter. Assume further that the gain factors $G_i(\nu)$ are rectangle ('window') functions of frequency, i.e. $G_i(\nu) = G_i\Pi(\nu/B)$ with $\Pi(\nu/B) = 1$ for $-B/2 < \nu < +B/2$ and 0 everywhere else. Moreover take $\Delta_i(\nu) = \Delta_i$ to be independent of frequency within the frequency bandwidth B of the system. Expression (2.13) can now be integrated over all frequencies to arrive at the total noise power output:

$$N_o(\nu) \;=\; \Delta k T_s G B, \text{ with } \Delta \;=\; \left[1 + \sum_{i=1}^{n} \frac{(\Delta_i - 1)}{\prod_{k=0}^{i-1} G_k}\right], \text{ in which } G_0 = 1 \qquad (2.14)$$

The expression for the noise factor Δ in (2.14) is called the *cascade rule* of Friis that quantifies the influence of successive processing stages on the **S/N** ratio. If we assume that in general the power gain factors of the individual stages $G_i \gg 1$ and that the individual noise factors of each stage Δ_i are comparable in magnitude we find $\Delta \simeq \Delta_1$.This shows that the first processing stage coupled to the sensor output largely determines the **S/N** ratio that can eventually be attained.

2.4 Noise equivalent circuit front-end interface

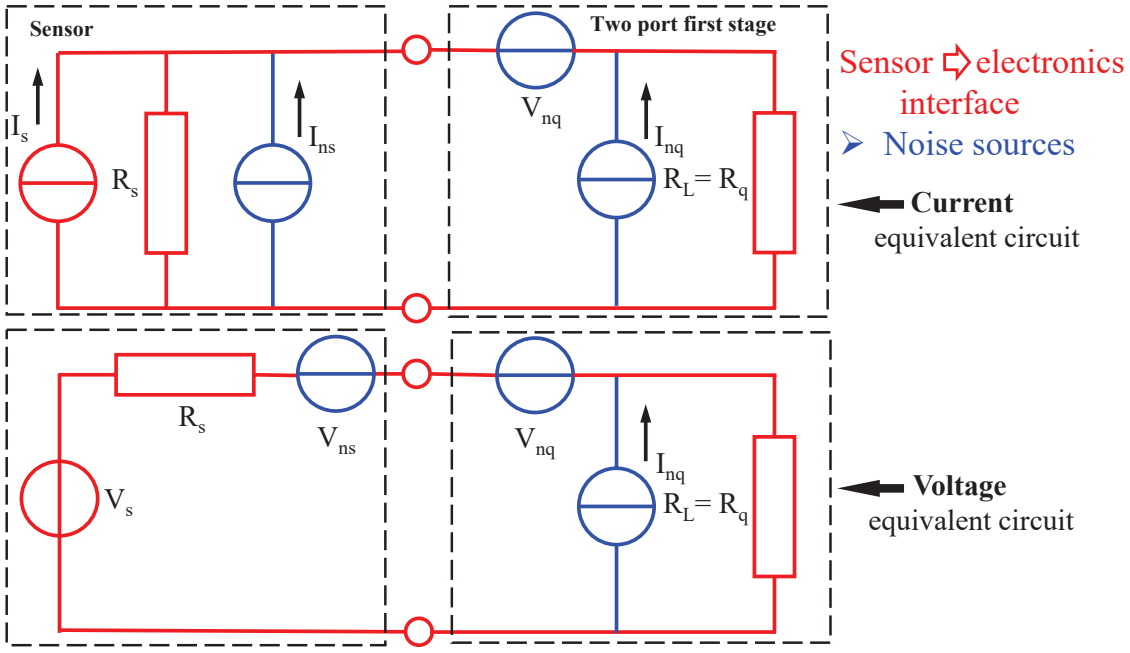


Figure 2.2: *Noise equivalent circuits for a sensor→ front-end interface. All signal sources and resistive loads in red, all noise equivalent sources in blue.*

Figure 2.2 shows the current and voltage noise equivalent circuits of the sensor interface with the front-end electronics stage. The sensor can be replaced by a noiseless resistor in series with an equivalent voltage noise source V_{ns} or parallel with an equivalent current

noise source I_{ns}. All noise sources in the front-end electronics stage are transformed back to the input as equivalent current and voltage sources I_{nq} and V_{nq}. The input impedance R_q of the two port is the load impedance for the sensor R_L. As shown earlier, maximum power extraction will occur for $R_L = R_s$, so we have available signal power from the sensor $V_s^2/4R_s$ and noise power from the sensor $\overline{V_{ns}^2}/4R_s$. The noise power contribution from the two port amounts to $\overline{V_{nq}^2}/4R_s + \overline{I_{nq}^2}R_s/4$. The power S/N-ratio at the sensor output can be expressed as the ratio of the signal power W_s to the noise power W_n:

$$\frac{\mathbf{S}}{\mathbf{N}} = \frac{W_s}{W_n} = \frac{V_s^2/4R_s}{\overline{V_{ns}^2}/4R_s + \overline{V_{nq}^2}/4R_s + \overline{I_{nq}^2}R_s/4} = \frac{V_s^2}{\overline{V_{ns}^2} + \overline{V_{nq}^2} + \overline{I_{nq}^2}R_s^2} \tag{2.15}$$

The noise contribution of the two port $W_{nq} = (\overline{V_{nq}^2}/4R_s) + (\overline{I_{nq}^2}R_s/4)$ can be minimized by an optimal choice for R_s. Evaluating $\partial W_{nq}/\partial R_s = 0$ we get $R_s = \sqrt{\overline{V_{nq}^2}}/\sqrt{\overline{I_{nq}^2}}$. In that case the maximum attainable S/N-ratio, using (2.15), becomes:

$$\left(\frac{\mathbf{S}}{\mathbf{N}}\right)_{max} = \frac{V_s^2}{\overline{V_{ns}^2} + 2\overline{V_{nq}^2}} = \frac{W_s}{W_{ns} + \frac{1}{2}\sqrt{\overline{V_{nq}^2}} \cdot \sqrt{\overline{I_{nq}^2}}} \tag{2.16}$$

The sensor total noise power W_{ns} needs to be as low as possible, this requires a thorough analysis of the frequency characteristics of the sensor signal and its various intrinsic noise sources, e.g. thermal noise, shot noise, $1/f$-noise and microphony, that must lead to an optimal choice of the transfer function $|H(2\pi j\nu)|^2$ for the sensor signal. Proper design of the input two port stage should minimize the $\sqrt{\overline{V_{nq}^2}} \cdot \sqrt{\overline{I_{nq}^2}}$ term in (2.16). This is strongly dependent on the choice of the input transistor stage and the feedback circuitry. For instance a JFET with low input capacitance is preferable over a MOSFET, use a common-emitter design, employ bipolar transistors.

Chapter 3

Sensor physics: conductivity

3.1 Energy band structures in solids, valence and conduction bands

Figure (3.1) shows, as an example, the electron configuration of the Germanium atom with the corresponding energy level diagram. This diagram is only valid for a single isolated atom. Beware: the energy levels of the orbital electrons with respect to the nucleus, i.e. inside the nucleus potential well, increase upward, the electrons tend to

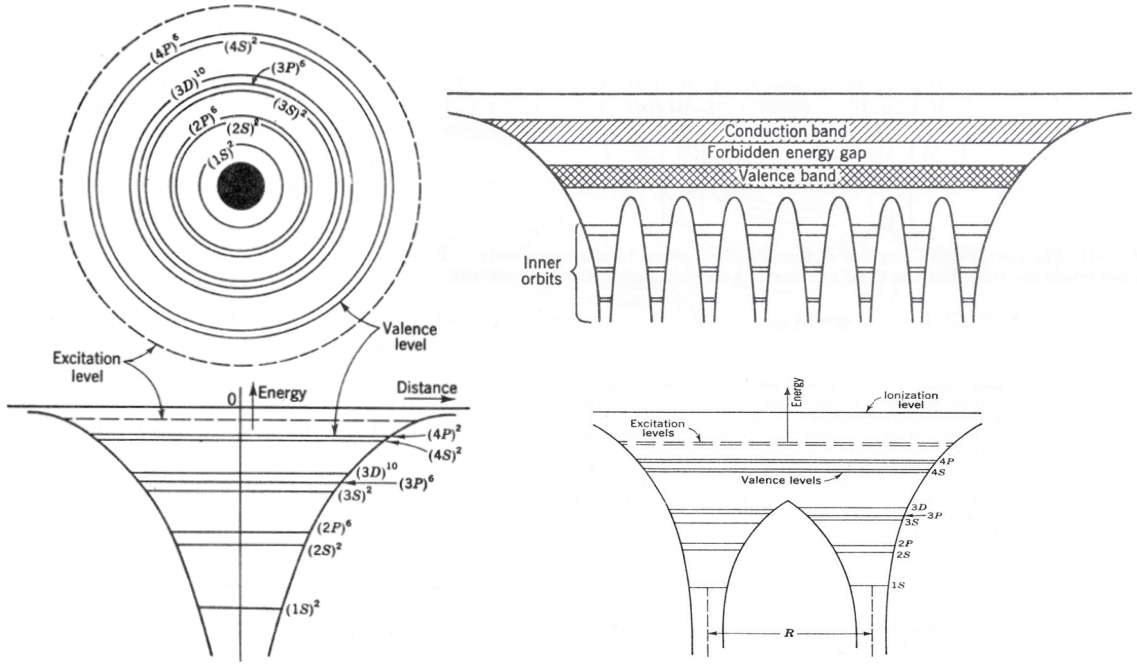


Figure 3.1: *Left: Orbital and energy level diagram of a single Germanium atom. Lower right: Energy level diagram of two atoms in close proximity, the energy levels are split. Upper right: Energy level diagram for a crystal lattice containing several atoms, the conduction-band and the valence-band energy levels extend throughout the crystal and are characterized by their volume Density of States (DoS).*

settle themselves at the lowest possible level in this potential well. The zero level potential, indicated as a reference in (3.1), corresponds to a free electron without any kinetic energy: the ionization potential. The electric potential is opposite to the electron potential, i.e. running more positive deeper in the potential well. In energy level diagrams as shown here, one almost exclusively indicates the electron energy or *electron potential*.
As a consequence of the fact that atoms in a solid have settled in a stable lattice structure, mutual influencing of energy levels between individual atoms will occur. This gives rise to high density splitting of the original levels resulting in energy band structures for the outermost orbital electrons. This implies that these electrons are no longer associated with individual atoms, but are collectively shared in the lattice structure. Despite the fact that the number of split-up energy levels is finite, their high density with respect to the electron potential can essentially be regarded as a continuous energy range.
To assess the conductivity of a solid two energy bands are relevant:

- Valence band: a group of energy levels that contain the valence electrons that are responsible for the inter-atomic connections. The valence electrons determine the *chemical* properties of a substance.

- Conduction band: the outermost energy band that either partly overlaps the valence band or that is completely separated from it by a forbidden zone by a so-called *energy gap*. The *electrical* properties of a solid are governed by the energetic position of the valence band relative to the conduction band, i.e. the degree of overlap or, alternatively, the width of the energy gap. So the mutual distance between the valence band and the conduction band determines whether a material is a good conductor, a semiconductor or an insulator. These three cases are shown in figure (3.2).

The energy level structure for Copper, an excellent conductor, shows that the valence band and the conduction band do show substantial overlap and are only partly filled with electrons (Cu contains only one 4S-electron). The energy bands for Silicon, a semiconductor, are separated by an energy gap of $\approx$ 1.1 eV. In this case electrons residing in the valence band will have to cross the energy gap to the conduction band (that was originally empty) to contribute to the conductance. This can, for example, be accomplished by thermal excitation. If the gap size increases to several eV, the probability for the valence electrons to cross the gap diminishes rapidly and materials with such wide gaps are classified as insulators that exhibit practically no electrical conductance. In figure (3.2) the example of diamond is shown that has a gap width of 7 eV.
The specific resistivity of materials can roughly be categorized as:

- Conductors: $\Rightarrow$ 10 milliΩ·cm
- Semiconductors: $\Rightarrow$ 1 MegaΩ·cm
- Insulators: $\geq$ 1 MegaΩ·cm

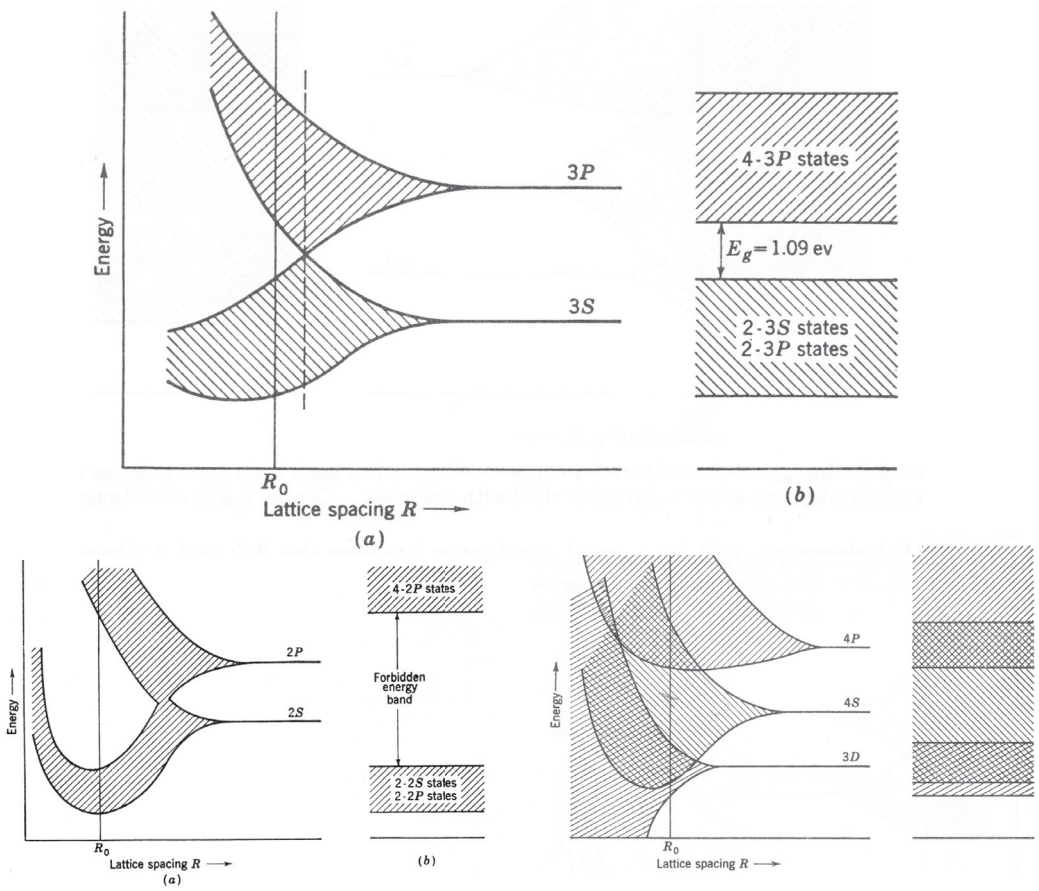


Figure 3.2: *(Lower left) The energy levels[bands] in the tetrahedral structure of diamond as a function of lattice spacing, the stable energy level of the diamond crystal is indicated at the lattice spacing R_0. The large forbidden energy gap characterizes diamond as an insulator. (Lower right) The energy levels[bands] structure for copper, the bands overlap and are only partly filled with electrons, hence copper is a very good conductor. (Upper picture) The energy levels of Silicon as a function of lattice spacing, the stable spacing R_0 for the silicon crystal is indicated leading to a relatively small energy gap of $\approx$ 1.1 eV. At room temperature sufficient electron-hole pairs are excited across the gap to allow appreciable conductivity, hence classification as semiconductor.*

3.2 Charge carrier distribution functions

3.2.1 Electron distribution function for metals

To calculate the energy distributions of the relevant charge carriers (electrons and holes), two distribution functions are required:

- The energy-volume density of the available energy levels in the valence band $g_v(E)$ and in the conduction band $g_c(E)$.
- The *average* occupation probability (*occupation number*) for an electron at a particular energy level $f(E)$ or for a hole $[1 - f(E)]$.

The function $f(E)$ is the *equilibrium* distribution law of Fermi-Dirac for fermions. The *average* occupation number per energy level ranges from 1 to 0, since fermions are subjected to the Pauli exclusion principle:

$$f_F(E) = \frac{1}{1 + e^{(E-E_F)/kT}} \tag{3.1}$$

In equation (3.1) E_F represents the *Fermi energy*. The Fermi energy, also called the Fermi level, is the electrochemical potential of electrons in a material and in this way it represents the averaged energy of electrons in the material.
At a temperature $T = 0$ we have: $E \leq E_F \Rightarrow f_F(E) = 1$ and for $E > E_F \Rightarrow f_F(E) = 0$, i.e. all the energy levels are occupied up until E_F and all energy levels above E_F are empty. This is shown in figure (3.3) at the left side for $T = 0$ along with $f_F(E)$ for a few more elevated temperatures.
For a metal the energy-volume density of the available energy levels in the conduction band can be shown to exhibit a parabolic energy dependence, i.e. $g(E) = A_m E^{\frac{1}{2}}$. This yields for the electron distribution function:

$$n_F(E) = g(E) f_F(E) = A_m \frac{E^{\frac{1}{2}}}{1 + e^{(E-E_F)/kT}} \tag{3.2}$$

We can compare this with the well-known Maxwell-Boltzman distribution of free moving independent particles in rarefied gases:

$$n_{MB}(E) = A_{MB}\, E^{\frac{1}{2}}\, e^{-E/kT} \tag{3.3}$$

Apparently these distributions will only become conformal, i.e. $n_F(E) \propto n_{MB}(E)$, in case $E - E_F \gg kT$ and, also, $E \gg E_F$.

3.2.2 Electron-hole distribution functions for intrinsic semiconductors

The position of the Fermi level in the energy level diagram governs the Fermi function, from figure (3.3) it is clear that for a semiconductor the Fermi level is located in between the lower boundary of the conduction band and the upper boundary of the valence band since for $T = 0$ the valence band is completely filled and the conduction band is completely empty (no conduction). If the temperature goes up, an increasing number of electrons will reach the conduction band owing to thermal agitation. These conduction band electrons can move as free charge carriers through the crystal lattice and determine the level of electrical conductivity of the semiconductor in question. In practise it is often allowed to work with the *effective* density of energy levels and charge carriers by referencing to the lower boundary E_c of the conduction band for electrons and to the upper boundary E_v of the valence band for holes. In this way the energy dependence of the density parameters can be eliminated, so by introducing these effective densities

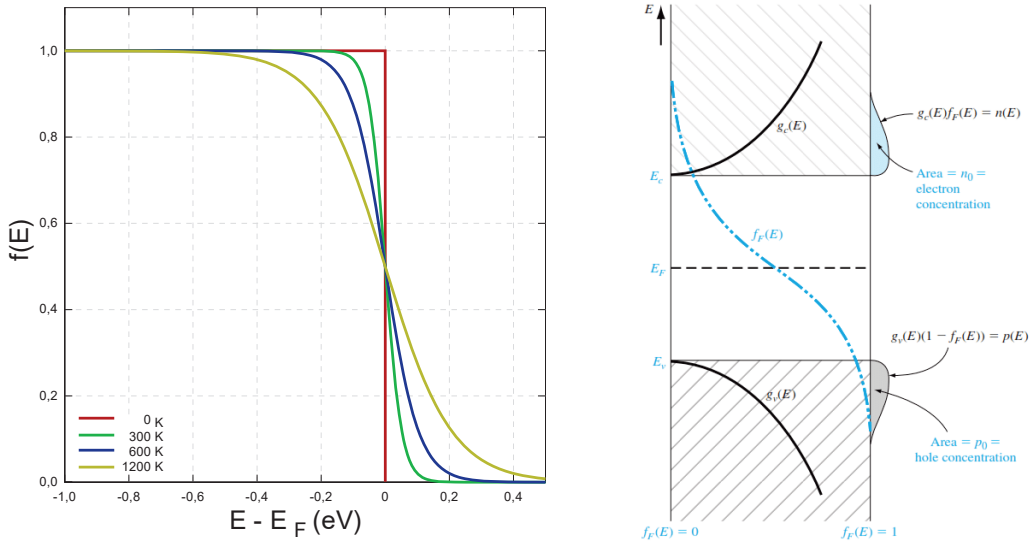


Figure 3.3: *(Left) The Fermi distribution function $f_F(E)$ displayed for different values of the absolute temperature T. (Right) The Fermi function at room temperature and the energy dependence of the volume density of states $g_{v,c}(E)$ in the valence and conduction bands. The electron (conduction band) and hole (valence band) concentrations at energy levels between E and $E + dE$ follow from $n(E)dE = g_c(E)f_F(E)dE$ and $p(E)dE = g_v[1 - f_F(E)]dE$ respectively. Expressions for the total concentrations $n_0 = p_0$ follow from integration over all energy levels, see text.*

we can write:

$$\begin{aligned} n_c &= g_c\, f_F(E_c) &&= \text{effective } \textit{total} \text{ electron volume density} \\ g_c &= \frac{n_c}{f_F(E_c)} &&= \text{effective } \textit{total} \text{ level volume density conduction band} \\ p_v &= g_v\,[1 - f_F(E_v)] &&= \text{effective } \textit{total} \text{ hole volume density} \\ g_v &= \frac{p_v}{[1 - f_F(E_v)]} &&= \text{effective } \textit{total} \text{ level volume density valence band} \end{aligned} \tag{3.4}$$

We can compute these effective density parameters analytically in good approximation by introducing the condition that $(E_c - E_F)$ and $(E_F - E_v)$ are both $\gg kT$, i.e. E_F lies several kT-values beneath E_c and, alternatively, E_F lies several kT-values above E_v. This assumption is certainly justified in the case of the well-known semiconductor materials Silicon and Germanium, since we have $kT \approx 26$ meV at $T = 300$ K, whereas the energy band gap for Si = 1106 meV and for Ge = 670 meV at $T = 300$ K. More specifically let's assume in the following derivation that $(E_c - E_F)$ and $(E_F - E_v)$ are both $\geq 3kT$.

We can now compute the total, energy independent, concentration (volume density) of the electrons in the conduction band and the holes in the valence band by integration over all relevant energy levels, i.e. the areas indicated in blue with n_0 and p_0 at the

right hand side in figure (3.3):

$$n_c = \int_{E_c}^{E_{top}} n(E)dE = \int_{E_c}^{E_{top}} g(E)\, e^{-(E-E_F)/kT} dE \quad \text{and} \tag{3.5}$$

$$p_v = \int_{E_{bottom}}^{E_v} p(E)dE = \int_{E_{bottom}}^{E_v} g(E)\, e^{(E-E_F)/kT} dE \tag{3.6}$$

where we have approximated the Fermi-Dirac distribution function by the substitutions $f_F(E) = e^{-(E-E_F)/kT}$ and $[(1 - f_F(E)] = e^{(E-E_F)/kT}$, which is justified given the energy level range over which the integration is to be performed.
From quantum mechanics we have an expression for the energy dependent density of the energy levels $g_c(E)$ with reference to the lower boundary of the conduction band E_c and $g_v(E)$ with reference to the upper boundary E_v of the valence band:

$$g_c(E) = 4\pi \left(\frac{2m_n^*}{h^2}\right)^{\frac{3}{2}} \sqrt{E - E_c} \quad \text{and} \quad g_v(E) = 4\pi \left(\frac{2m_p^*}{h^2}\right)^{\frac{3}{2}} \sqrt{E_v - E} \tag{3.7}$$

The mass m^* in these expressions refers to the effective mass of the charge carrier, which depends on the specific location and on the lattice structure. For Germanium we have $m_n^*/m_n = 0.55$ and $m_p^*/m_p = 0.35$, for Silicon $m_n^*/m_n = 1.08$ and $m_p^*/m_p = 0.70$ respectively. In both cases the value of m_n^* is larger than m_p^*, for the semiconductor GaAs it is the other way around.
Substitution of (3.7) in (3.6) and a change of variables yields:

$$n_c = 4\pi \left(\frac{2m_n^*}{h^2}\right)^{\frac{3}{2}} e^{-(E_c-E_F)/kT} \int_0^{\infty} \sqrt{x}\, e^{-x/kT} dx \quad \text{with} \quad x = E - E_c \tag{3.8}$$

$$p_v = 4\pi \left(\frac{2m_n^*}{h^2}\right)^{\frac{3}{2}} e^{-(E_F-E_v)/kT} \int_0^{\infty} \sqrt{x}\, e^{-x/kT} dx \quad \text{with} \quad x = E_v - E \tag{3.9}$$

Note: the integration boundaries E_{top} and E_{bottom} can be extended to ∞ and $-\infty$ respectively, since the integrand is zero beyond these values, the sign and boundary interchange in case of p_v follows from the change of variables with $dx = -dE$.
Performing the integration, using $\int_0^{\infty} \sqrt{x}\, e^{-nx} dx = 1/(2n)\sqrt{\pi/n}$, we get:

$$n_c = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{\frac{3}{2}} f_F(E_c) \qquad \Rightarrow g_c = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{\frac{3}{2}} = A_c \cdot T^{\frac{3}{2}} \tag{3.10}$$

$$p_v = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{\frac{3}{2}} [1 - f_F(E_v)] \Rightarrow g_v = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{\frac{3}{2}} = A_v \cdot T^{\frac{3}{2}} \tag{3.11}$$

Regarding the numerical values of the constants we have for Germanium $A_c(Ge) = 1.97 \cdot 10^{21}$ m^{-3}, $A_v(Ge) = 1.00 \cdot 10^{21}$ m^{-3} and for Silicon $A_c(Si) = 5.42 \cdot 10^{21}$ m^{-3},

$A_v(Si) = 2.83{\cdot}10^{21}\ \mathrm{m}^{-3}$. The differences in these constants for the conduction band and the valence band is entirely due to the difference in effective mass between the electrons and holes in both materials, in case $m_n^* = m_p^*$ we would have obtained $g_c = g_v$.
Since the electrons and holes are produced in pairs, we should have by definition $n_c = p_v$. From this equality we can derive the value of the Fermi energy:

$$E_F = \frac{E_c + E_v}{2} + \frac{3}{4}kT \ln \frac{m_p^*}{m_n^*} \quad \Rightarrow \quad E_F = E_v + \frac{E_g}{2} + \frac{3}{4}kT \ln \frac{m_p^*}{m_n^*} \tag{3.12}$$

in which E_g is the width of the energy gap. Hence the Fermi level lies very close to the midgap $[(E_c + E_v)/2]$, a small difference is caused by the difference in effective masses (= effective densities) in the valence and the conduction band that introduces a slight temperature dependence of the Fermi level.
We can also write the electron/hole volume density as a function of temperature:

$$n_c = A_c\, T^{\frac{3}{2}}\, e^{-(E_c - E_F)/kT} = n_i \tag{3.13}$$

$$p_v = A_v\, T^{\frac{3}{2}}\, e^{-(E_F - E_v)/kT} = n_i \tag{3.14}$$

in which $n_i = n_c = p_v$ signifies the density of free charge carriers in the *intrinsic* semiconductor material (index i). At $T = 300$ K we have for Germanium $n_i \approx 1.7 \cdot 10^{19}$ m^{-3} and for Silicon $n_i \approx 10^{16}$ m^{-3}.
The electron-hole recombination speed is evidently proportional to the product of the electron and hole concentrations, hence we have a probability:

$$np = n_i^2 = A_c A_v\, T^3\, e^{-(E_c - E_v)/kT} = A_c A_v\, T^3\, e^{-E_g/kT} \tag{3.15}$$

The recombination rate is clearly dependent on the gap energy and the temperature, but does not depend on the actual position of the Fermi level.
The *conductivity* of an intrinsic semiconductor can be assessed starting from the expression for the current density $\vec{j}$:

$$\begin{aligned}
\vec{j} &= -n_i q \vec{v}_n + n_i q \vec{v}_p, \text{ with } \vec{v}_n/\vec{v}_p \text{ the drift velocity of electrons/holes respectively} \\
\vec{j} &= \sigma \vec{E}, \ \sigma = \text{ the material conductivity} \\
\vec{v}_n &= -\mu_n \vec{E}, \ \mu_n = \text{mobility of electrons and } \vec{v}_p = \mu_p \vec{E}, \ \mu_p = \text{mobility of holes } \Rightarrow \\
\sigma &= n_i q(\mu_n + \mu_p) \quad \text{with substitution of } n_i \quad \Rightarrow \\
\sigma &= (A_c A_v)^{\frac{1}{2}}\, T^{\frac{3}{2}} q(\mu_n + \mu_p)\, e^{-(E_g/2kT)}
\end{aligned} \tag{3.16}$$

The resistivity ρ is the inverse of the conductance σ, i.e. $\rho = 1/\sigma$. The factor $T^{\frac{3}{2}}$ and the mobilities change relatively slowly with the temperature compared to the exponential term, hence the logarithm of ρ varies nearly linearly with $1/T$ as shown in figure (3.4). So we can write:

$$\ln \rho = -\ln \sigma \propto \frac{1}{T} \quad \text{i.e. a negative temperature coefficient} \tag{3.17}$$

This proportionality shows that from the slope of a plot as displayed in figure (3.4) one can determine the value of the energy gap E_g.

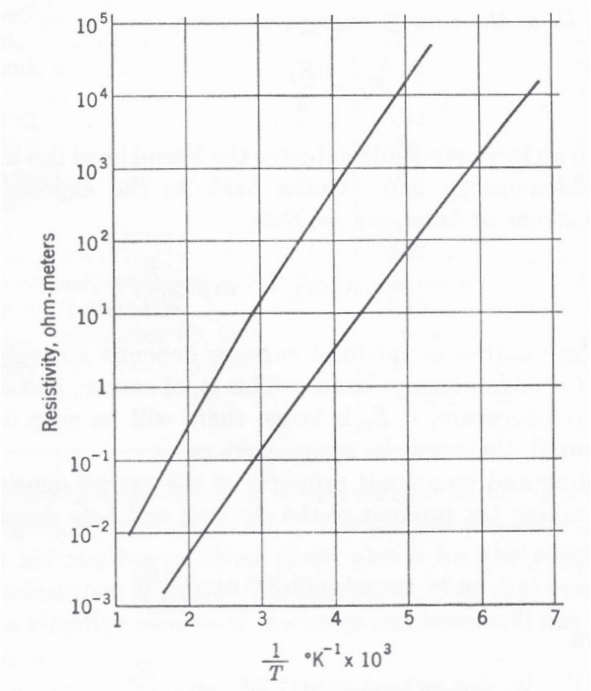


Figure 3.4: *Temperature-resistance curves of two typical intrinsic semiconductors. From the slope of the curves one finds the energy gaps to be about 0.63 and 0.53 eV.*

3.2.3 Extrinsic semiconductors

Extrinsic semiconductors are produced by injection of impurities in the crystal lattice. In case the impurities constitute a *pentavalent* element, i.e. an element with five valence electrons, this will result in weakly bound electrons. These impurities are called donor elements, examples are Arsenicum (As), Phosphor (P), Antimony (Sb) and Bismuth (Bi). In case the impurities constitute an *trivalent* element, i.e. an element with three valence electrons, this will result in weakly bound holes. These elements are called acceptor elements, examples are Boron (B), Aluminum (Al), Gallium (Ga) and Indium (In). Their typical binding energy in impurity doped Germanium is of the order 10-12 meV, in impurity doped Silicon the binding energy is typically 50 meV. The impurity energy levels and the associated Fermi levels are shown in figure (3.5).

If an electron resides on the donor energy level and if no electron resides on the acceptor level, we designate the donor and acceptor level as *neutral*. If the donor electrons reside in the conduction band or the holes reside in the valence band due to electron capture, we designate the donor and acceptor level as *ionized*. In material that is doped with donor impurities the electrons are the so-called *majority* charge carriers and govern the conductivity in the conduction band of a so-called *n-type* semiconductor [red in figure (3.5)], the holes are in this case the *minority* charge carriers. In material that is doped with acceptor impurities, the holes are the majority charge carriers and govern the conductivity in the valence band of a so-called *p-type* semiconductor [blue in figure (3.5)], the electrons constitute in this case the minority charge carriers.

To assess the conductivity of doped semiconductor material we need to determine the

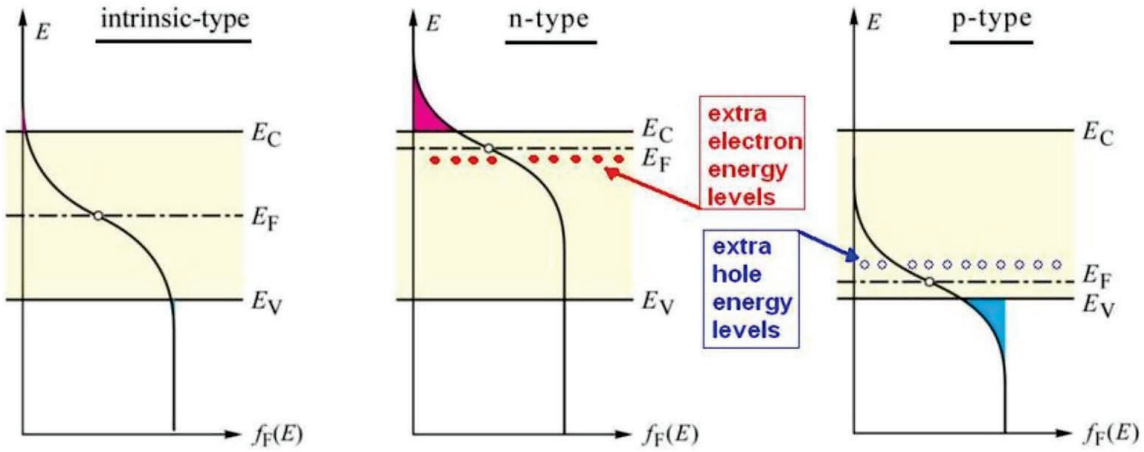


Figure 3.5: *Shift of the Fermi level in doped semiconductors. The donor (E_d) and acceptor (E_a) energy levels are also indicated.*

volume density of electrons and holes in the conduction band and the valence band respectively. This can be calculated if we know the position of the Fermi level. The calculation of this position is rather complicated if we have to account for both the intrinsic and the extrinsic charge carriers. However in most practical cases this is not really necessary. To underpin this, let us start from the following relations, employing the effective density approach:

$$n_c = N_c f_F(E_c) \quad \text{density of electrons in conduction band} \tag{3.18}$$
$$p_v = N_v[1 - f_F(E_v)] \quad \text{density of holes in valence band} \tag{3.19}$$
$$n_d = N_d f_F(E_d) \quad \text{density of occupied donors} \tag{3.20}$$
$$n_a = N_a f_F(E_a) \quad \text{density of occupied acceptors} \tag{3.21}$$
$$p_v + (N_d - n_d) = n_c + n_a \quad \text{charge neutrality + charge = − charge} \tag{3.22}$$

Next we may simplify the calculation by setting N_d or N_a equal to zero, i.e. we have either only donor impurities or acceptor impurities.
As an example let's assume that we have no acceptor impurities, i.e. $N_a = 0$, moreover assume that $p_v \approx 0$ implying that the conductance is solely due to donor injected electrons: $n_c = N_d - n_d$ from the charge neutrality equation. This means that $p_v \ll n_c$, an assumption that we now need to verify! For this purpose let us take a relatively low donor concentration of $N_d = 10^{22}$ m^{-3} and a band gap between the donor level E_d and the lower boundary of the conduction band E_c of ≈ 10 meV (typical for Germanium). Assuming $(E_c - E_F) \gg kT (> 3kT$ as we used before) and $T = 300K$ we can write:

$$\begin{aligned} n_c &= A_c T^{\frac{3}{2}} e^{-(E_c - E_F)/kT} \quad \Rightarrow \\ N_d - n_d = N_d[1 - f_F(E_d)] = N_d[1 - f_F(E_c - 0.01)] &= N_d[1 - e^{-(E_c - 0.01 - E_F)/kT}] \Rightarrow \\ 1.97 \cdot 10^{21}(300)^{\frac{3}{2}} e^{-(E_c - E_F)/kT} &= 10^{22}[1 - e^{-(E_c - 0.01 - E_F)/kT}] \end{aligned}$$

Solving for $(E_c - E_F)$ yields: $E_c - E_F = 179$ meV, this is substantially closer to the conduction band than in the case of an intrinsic Germanium crystal ($E_c - E_F \approx 335$

meV). Knowing now the position of the Fermi level, we can check the hole density in the valence band:

$$p_v = 1.00 \cdot 10^{21}\, T^{\frac{3}{2}} e^{-(E_F - E_v)/kT} = 1.00 \cdot 10^{21}\, (300)^{\frac{3}{2}}\, e^{-(0.670-0.179)/kT} \approx 2.8 \cdot 10^{16} \mathrm{m}^{-3}$$

This result indeed confirms that $p_v \ll n_c$, since n_c is of the order of 10^{22} m^{-3} owing to the fact that at 300 K practically all the donors are ionized, i.e. $kT \gg E_d$. In the *extrinsic limit*, when either the donor level (n-type) or the acceptor level (p-type) is completely ionized, we have $n_c = N_d$ or $p_v = N_a$ and, consequently, the conductivity can be expressed as:

$$\text{For n-type semiconductors } \sigma = N_d\, q\, \mu_n \tag{3.23}$$

$$\text{For p-type semiconductors } \sigma = N_a\, q\, \mu_p \tag{3.24}$$

In this situation the position of the Fermi level can be straightforwardly calculated with the aid of equations (3.10) and (3.11) for n-type ($n_c = N_d$) and p-type ($p_v = N_a$) semiconductors respectively:

$$E_c - E_F = kT \ln \frac{g_c}{N_d} \quad \text{for n-type and} \quad E_F - E_v = kT \ln \frac{g_v}{N_a} \quad \text{for p-type} \tag{3.25}$$

If the ambient temperature T is substantially raised, the contribution of thermally excited electron-hole pairs becomes increasingly important, the extrinsic character of the semiconductor becomes more and more intrinsic, the Fermi level decreases ultimately again to the midgap value for intrinsic material. We can also readily compute the position of the Fermi level in the low temperature limit. Consider:

$$\begin{aligned} n_c &= N_c f_F(E_c) \text{ and} \\ n_c &= N_d - n_d = N_d[1 - f_F(E_d)] \quad \Rightarrow \\ N_c f_F(E_c) &= N_d[1 - f_F(E_d)] \end{aligned}$$

If the temperature is sufficiently low so that $E_C - E_F \gg kT$ and $E_F - E_d \gg kT$, we may write:

$$N_c\, e^{-(E_c - E_F)/kT} = N_d\, e^{(E_d - E_F)/kT} \quad \text{solving for } E_F \text{ we get} \Rightarrow$$

$$E_F = \frac{E_c + E_d}{2} - \frac{kT}{2} \ln \frac{N_c}{N_d} \quad \Rightarrow \frac{E_c + E_d}{2} \text{ for } T \Rightarrow 0 \tag{3.26}$$

Hence, in the low temperature limit the Fermi level sits halfway between the donor level and the lower boundary of the conduction band. This evolution of the Fermi level position as a function of temperature is displayed in figure (3.6). From this picture we can make the following observations:

- At low temperature the Fermi level resides midway between the filled donor level and the lower boundary of the conduction band. With increasing temperature the Fermi level migrates ultimately towards a midgap energy position between the conduction band and the valence band when the intrinsic limit has been reached.

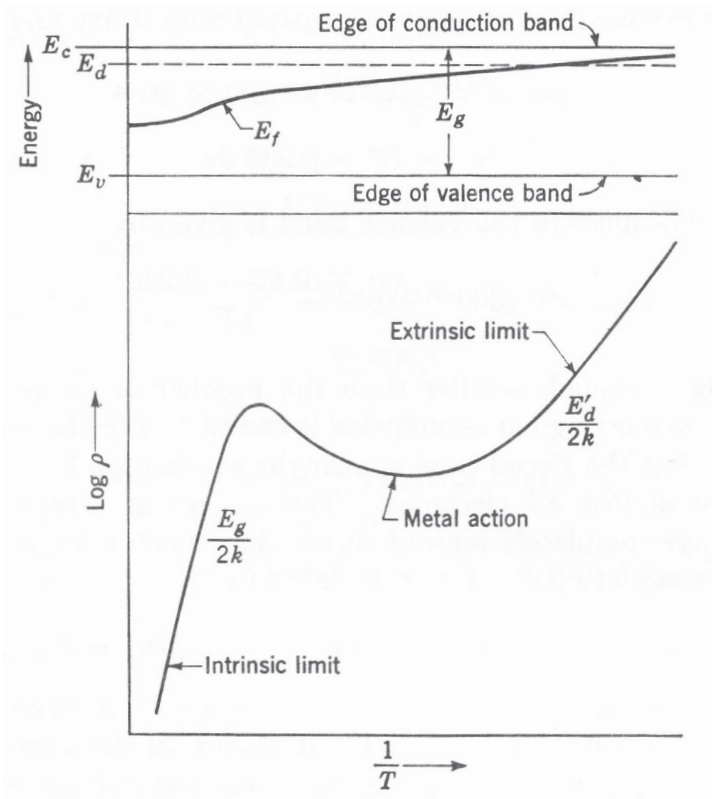


Figure 3.6: *The variation of the Fermi level with temperature. The corresponding variation of the resistivity is also shown for comparison. Note the limiting cases for intrinsic and extrinsic conditions.*

- There are two regions with a negative temperature coefficient: the intrinsic limit and the extrinsic limit.

- In between this limiting behavior there exists a region with a positive temperature coefficient, i.e. the material exhibits a metal characteristic. This behavior follows from the decrease in mobility of the charge carriers at more elevated temperatures, whereas the donor (acceptor) levels are already fully ionized but the enhanced thermal excitation of electron-hole pairs has not set in yet since the temperature is still too low. This temperature region evidently represents a transition phase.

We noted earlier that the the electron hole recombination rate $p_v n_c = n_i^2 = f(E_g, T)$ does depend on the width of the band gap and the temperature, but is independent of the position of the Fermi level. If the recombination mechanism is independent of the carrier density than the above recombination rate equation will always hold. If we have a material with both donor and acceptor impurities that are all fully ionized, the charge neutrality equation dictates $n_c + N_a = p_v + N_d$ and with $p_v n_c = n_i^2$ we can solve

for the charge carrier concentrations:

$$n_c = \frac{(N_d - N_a)}{2} + \left[n_i^2 + \left(\frac{N_d - N_a}{2} \right)^2 \right]^{\frac{1}{2}} \tag{3.27}$$

$$p_v = \frac{(N_a - N_d)}{2} + \left[n_i^2 + \left(\frac{N_a - N_d}{2} \right)^2 \right]^{\frac{1}{2}} \tag{3.28}$$

If $N_a = N_d$, we have again $p_v = n_c = n_i$. This case is labeled as *complete compensation* of donor and acceptor impurities. In practice two extreme cases are of importance:

- Pure n-type material with $\Rightarrow N_d \gg n_i, N_a = 0$. The equilibrium concentration of holes then equals $p_v = n_i^2/N_d$.
- Pure p-type material $\Rightarrow N_a \gg n_i, N_d = 0$. The equilibrium concentration of electrons then equals $n_c = n_i^2/N_a$.

This phenomenon is called *minority carrier suppression*: an increase of the doping concentration will be accompanied by a proportional decrease in concentration of the minority charge carriers.
Example: compute the equilibrium concentrations n_c and p_v in doped Silicon at a temperature of 300 K containing $8 \cdot 10^{22}$ m^{-3} As-impurities and $2 \cdot 10^{22}$ m^{-3} B-impurities. For Si we have $n_i \approx 10^{16}$ m^{-3}. Applying equation (3.27) and noting that $n_i \ll (N_d - N_a)$ this results in $n_c = (N_d - N_a) = 6 \cdot 10^{22}$ m^{-3} in case of complete ionization. Solving equation (3.28) for $n_i \ll (N_d - N_a)$ we get $p_v = n_i^2/(N_d - N_a) = n_i^2/n_c$, we find $p_v = 1.7 \cdot 10^9$ m^{-3}. This shows a suppression of p_v by several orders of magnitude as compared to the intrinsic value, a clear demonstration of the magnitude of minority carrier suppression.
The Fermi level E_{F_i} of an intrinsic semiconductor is often used as a reference level for characterizing extrinsic semiconductors. From equations (3.10) and (3.11) we have for the intrinsic electron and hole concentrations:

$$n_c = p_v = n_i = g_c \, e^{-(E_c - E_{F_i})/kT} = g_v \, e^{-(E_{F_i} - E_v)/kT} \tag{3.29}$$

The extrinsic electron (n) and hole (p) concentrations may now be expressed in terms of the intrinsic carrier concentrations and the intrinsic Fermi level E_F:

$$n = n_i \, e^{(E_{F_e} - E_{F_i})/kT} \quad \text{and} \quad p = n_i \, e^{(E_{F_i} - E_{F_e})/kT} \tag{3.30}$$

Thus the energy separation from the Fermi level E_{F_e} to the intrinsic Fermi level E_{F_i} is a measure of the departure of the semiconductor from intrinsic material. Since E_{F_e} is above E_{F_i} in a n-type semiconductor we have $n > n_i > p$, as we found before.
Taking the above example with $n_c = 6 \cdot 10^{22}$ m^{-3} and $n_i = 10^{16}$ m^{-3}, we can now compute the position of E_{F_e} from $E_{F_e} - E_{F_i} = kT \ln n_c/n_i = 25.8 \ln (6 \cdot 10^{22}/10^{16})$, yielding $E_{F_e} = 402.7$ meV above the 'midgap' intrinsic Fermi level E_{F_i}. *Note*: We could also have calculated the position of the extrinsic Fermi level relative to the conduction band or the valence band by applying equation (3.25), e.g. $E_c - E_F = kT \ln g_c/(N_d - N_a) =$

$25.8 \ln (2.8 \cdot 10^{25}/6 \cdot 10^{22}) = 158.6$ meV. The sum of these two energies (561.3 meV) should obviously equal half the band gap of Si (553 meV) + the small temperature correction as given in equation (3.12). This temperature correction at 300 K amounts to -8.4 meV if the upper level of the valence band is taken as the reference, i.e. the intrinsic E_F is situated 544.6 meV above the upper boundary of the valence band and 561.4 below the lower boundary of the conduction band, in excellent agreement with the above result of 561.3 meV.

Chapter 4

Sensor physics: contact potentials

4.1 Gibbs free energy: thermodynamic potential

Thermodynamic equilibrium in matter requires that the so-called *thermodynamic potential* is everywhere the same in the material under consideration. In solid state physics we encounter in most cases conditions where temperature and pressure are kept constant. Application of the second law of thermodynamics then yields:

$$\begin{aligned} TdS &\geq dE + pdV \quad \text{equality in the case of TD-equilibrium} \quad \Rightarrow \\ dG &= d(E + pV - TS) \leq 0, \text{ with } p, T \text{ constant,} \\ G &= \text{Gibbs free energy or thermodynamic potential} \end{aligned} \tag{4.1}$$

For thermodynamic equilibrium G needs to be at its minimum value $\Rightarrow dG = 0$. Since pdV at atmospheric pressure is normally small as compared to dE and TdS, we have in that case:

$$dF = 0, \text{ with } F = E - TS \quad \text{(Helmholtz free energy)} \tag{4.2}$$

From solid state physics it can be shown that the Fermi energy, except for a constant, is equal to the free energy or thermodynamic potential. As a consequence, for equilibrium the Fermi level E_F needs to be the same everywhere in the specimen under consideration. Alternatively: no currents are allowed to run anywhere. To arrive at a minimum for G or F the entropy S needs to be maximized.

4.2 Metal-metal contact: the Volta effect

The potential energy of electrons inside a metal is lower than outside the metal (vacuum). The jump in potential (U_M) necessary to let the electrons exit starting from the Fermi level, is called the exit-potential or *work function* of the metal:

$$U_M = \phi_v - \frac{E_F}{q}, \ \phi_v = \text{vacuum potential and } q = \text{elementary charge}=1.602\cdot 10^{-19} \text{ [C]} \tag{4.3}$$

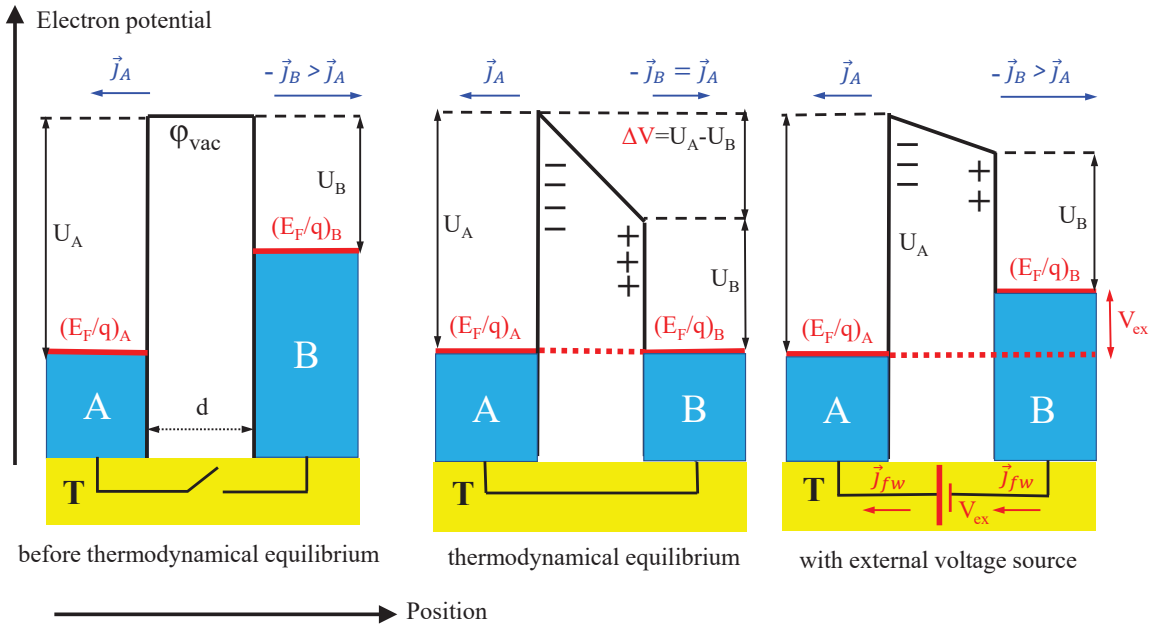


Figure 4.1: *Electron potential diagrams of a metal contact with different Fermi levels.(Left): Potentials* (U, E_F/q, V_e) *and current densities* ($\vec{j}$, *opposite to the direction of the electron flow!) before thermodynamic equilibrium. The metal contact is surrounded by a thermal bath maintaining the contact on a uniform temperature* T. *Open switch: no electrical contact.(Middle): Potentials and current densities after thermodynamic equilibrium has been established by making an electric connection.(Right): Potentials and current densities after equilibration when an external voltage source has been inserted in the electrical connection.*

If thermally excited electrons exit the metal and get caught in the suction of an externally applied electric field, an emission current density is created. The *magnitude* of this current density can be expressed by Richardson's formula for thermionic emission:

$$|j| = C \cdot T^2 \, e^{-qU_M/kT} \tag{4.4}$$

Consider now two metal slabs A and B with surfaces closely opposite to each other, in terms of exit potentials let's assume $U_A > U_B$. Next, both slabs are electrically connected. In case the potential between the slabs would remain unaltered, an emission current density $j_B > j_A$ would follow, giving rise to a permanent current. This is obviously in conflict with the second law of thermodynamics, which imposes that in equilibrium we need to have $E_{F_A} = E_{F_B} = E_F$. The situation is depicted in figure (4.1) that shows the transition from non-equilibrium (left picture) to equilibrium (middle picture). In case the Fermi electron potential E_F of metal B is higher than that of metal A, electrons will move from B to A and will built up a negative surface charge on A and leave behind a positive surface charge on metal B. At equilibrium, a potential drop $\Delta V = U_A - U_B$ across the vacuum gap will have developed, causing the (opposite) current densities from A and B to equalize on both sides and cancel, so that no net current will flow.

The current densities at equilibration become:

$$-\vec{j}_A \;=\; -\vec{j}_0\, e^{-qU_A/kT} \;=\; \vec{j}_0\, e^{-q(U_B+\Delta V)/kT)} \;=\; \vec{j}_B \quad \text{with } |\vec{j}_0| \;=\; C\cdot T^2 \tag{4.5}$$

The potential jump (contact potential) that settles the equilibration is called the *Volta effect* and we have $-\vec{j}_A = \vec{j}_B = \vec{j}_{eq}$. If we now apply a negative potential to metal B with an external voltage source V_{ex}, potential U_A and hence $\vec{j}_A$ will remain unchanged, however the electron potential on metal B is raised and the potential barrier seen from B to A has been lowered so that $\vec{j}_B$ will increase [see picture at the right in figure (4.1)]. The net current density $\vec{j}_{fw}$ flowing from A to B (electrons flowing from B to A) can now be expressed as:

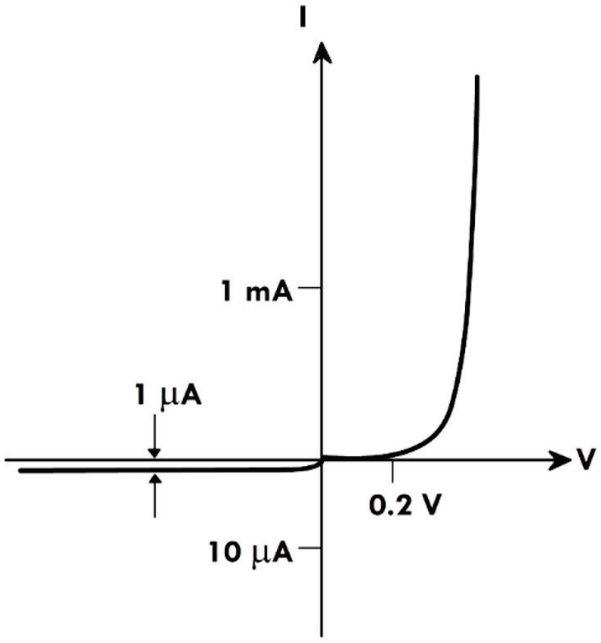


Figure 4.2: *Current-Voltage (I-V) diagram for a metal-metal junction. The I-V relation constitutes exponential functions as shown in equation (4.6) in case of a forward bias and in equation (4.7) in case of a reverse bias. The rectifying character of the characteristic is evident.*

$$\begin{aligned}\vec{j}_{fw} \;&=\; \vec{j}_B - \vec{j}_A \;=\; \vec{j}_0\left[e^{-q(U_B+\Delta V-V_{ex})/kT} - e^{-qU_A/kT}\right] = \\ &=\; \vec{j}_0\cdot e^{-qU_A/kT}\left[e^{qV_{ex}/kT}-1\right] \;=\; \vec{j}_{eq}\left[e^{qV_{ex}/kT}-1\right]\end{aligned} \tag{4.6}$$

If we, alternatively, apply a positive potential to metal B (*reverse bias*), the electron potential will be lowered and the potential barrier seen from B to A has been raised so that $\vec{j}_B$ will decrease. The net current density $\vec{j}_r$ is now reversed and flows from B to A (electrons flowing from A to B):

$$\begin{aligned}\vec{j}_r \;&=\; \vec{j}_B - \vec{j}_A \;=\; \vec{j}_0\left[e^{-q(U_B+\Delta V+V_{ex})/kT} - e^{-qU_A/kT}\right] = \\ &=\; -\vec{j}_0\cdot e^{-qU_A/kT}\left[1-e^{-qV_{ex}/kT}\right] \;=\; -\vec{j}_{eq}\left[1-e^{-qV_{ex}/kT}\right]\end{aligned} \tag{4.7}$$

Thus, the current-voltage characteristic has the shape of a rectifier.

4.3 Metal-semiconductor contact: the Schottky junction

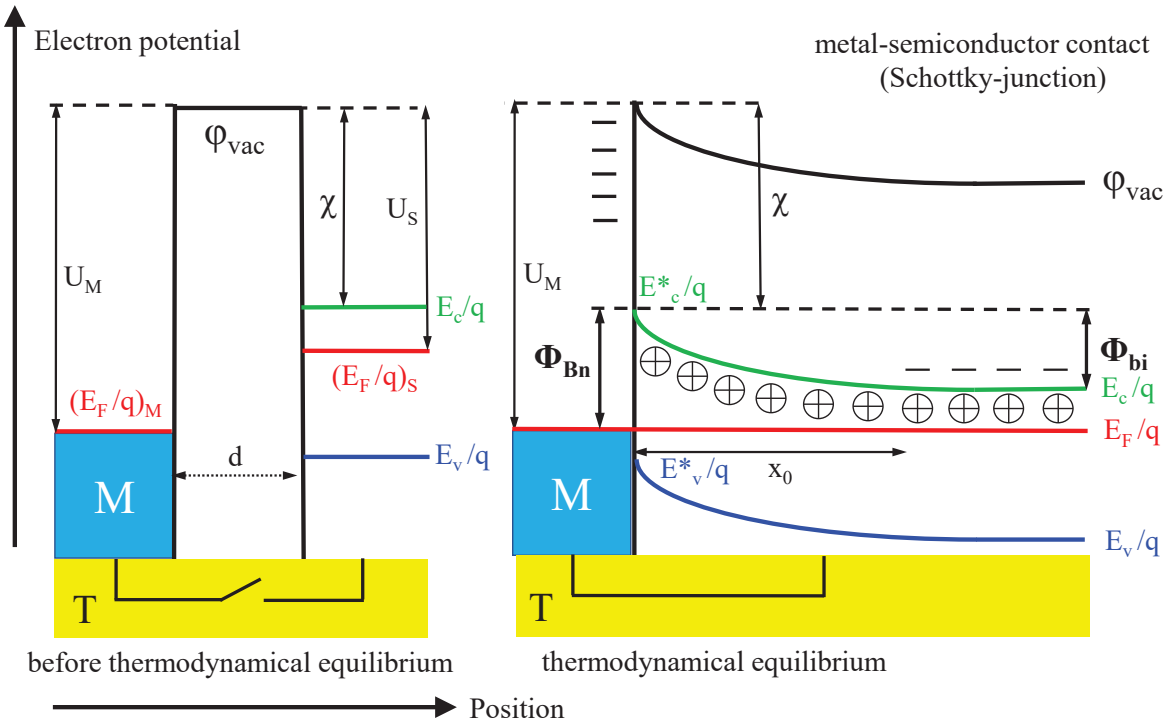


Figure 4.3: *Electron potential diagrams of a metal-semiconductor contact (Schottky junction) with different Fermi levels.(Left): Potentials* $(U, E_F/q, E_c/q, E_v/q, \chi)$ *before thermodynamic equilibrium. The contact is surrounded by a thermal bath at a uniform temperature* T*. Open switch: no electrical contact. (Right): Potentials after thermodynamic equilibrium have settled after electric connection. In case of a n-type semiconductor dominated by electron conduction, equalization of the Fermi levels, i.e. no net current through the junction, is established by the development of a negative surface charge on the metal surface and a positive space charge originating from the ionized donor atoms in the crystal lattice. The space charge region extends into the semiconductor, forming a depletion layer that is void of charge carriers.*

The electron potentials of a metal semiconductor contact, also known as a Schottky-junction, are displayed in figure (4.3). The picture on the left side shows the situation before thermodynamical equilibrium when no electrical contact has yet been established. We have the following potentials:

- U_M = the exit potential (work function) of the metal to vacuum.
- U_S = the exit potential (work function) of an n-type semiconductor to vacuum.
- χ = the *electron affinity* of the semiconductor: this is the exit potential the electrons in the conduction band need to bridge to reach vacuum. *Note*: the electron affinity can have a positive as well as a negative value.

For thermodynamical equilibrium the different Fermi levels in the metal and the semiconductor need to equalize. Assuming that $(E_F/q)_S > (E_F/q)_M$, electrons from the n-type semiconductor will be transported to the metal surface. This will result in:

- A negative surface charge at the metal side
- A positive space charge at the semiconductor side, resulting from depletion of the boundary layer leaving positively charged donor ions.

With $d = 0$, and neglecting surface effects, the following equilibrium situation will emerge, as can be seen from the band diagram in figure (4.3):

$$\Phi_{B_n} = U_M - \chi = \frac{(E_c^* - E_F)}{q} \tag{4.8}$$

$$\Phi_{bi} = U_M - U_S = \frac{[(E_F)_S - (E_F)_M]}{q} \tag{4.9}$$

$$\Phi_{B_n} = \Phi_{bi} + \frac{(E_c - E_F)}{q} \quad \text{with} \quad (E_c - E_F) = kT \ln\left(\frac{g_c}{N_d}\right) \tag{4.10}$$

In the above expressions Φ_{B_n} signifies the potential barrier for electrons (n) as seen from the metal to the semiconductor, this is commonly called the *Schottky barrier*. Φ_{bi} represents the potential barrier for conduction band electrons (n) as seen from the semiconductor to the metal, this is referred to as the built-in (bi) potential of the junction. Equation (4.10) gives their interrelation, where we inserted the expression for $(E_c - E_F)$ that was derived in equation (3.25). Values of Φ_{B_n} measured for some current metal-Si^n junctions are: 0.9 V (Pt/Si), 0.8 V (Au/Si), 0.5 V (Ti/Si), 0.6 V ($TiSi_2$/Si) and 0.75 V ($PdSi_2$/Si).
In contrast to the metal, where the negative charge is accumulated solely on the surface (no depletion layer), inside the semiconductor a space charge is formed that streches out over a certain depth x_0, the depletion layer that is almost void of any free charges. The positive charge is embodied in the ionized donor atoms in the n-type semiconductor material, fixed in the crystal lattice with a finite density dictated by the doping concentration. An estimate of the extent x_0 of the depletion layer can be obtained by solving the Poisson equation that relates the potential to the space charge density:

$$\frac{d^2\phi(x)}{dx^2} = -\frac{\rho}{\epsilon_0 \epsilon_r} \quad \text{with } \rho \text{ the space charge density} \tag{4.11}$$

For the depleted n-type region containing positively charged donor ions with $\rho = \rho_+ =$

$+qN_d$ we have:

$$\begin{aligned}
\frac{d^2\phi(x)}{dx^2} &= -\frac{qN_d}{\epsilon_0\epsilon_r} \Rightarrow \text{integration} \Rightarrow \frac{d\phi(x)}{dx} = -\frac{qN_d x}{\epsilon_0\epsilon_r} + C_1 = -E(x), \\
x &= x_0 \rightarrow E(x_0) = 0 \Rightarrow C_1 = \frac{qN_d x_0}{\epsilon_0\epsilon_r} \Rightarrow \\
E(x) &= \frac{qN_d}{\epsilon_0\epsilon_r}(x - x_0) \Rightarrow \phi(x) = -\int_+ E(x)dx = -\frac{qN_d}{2\epsilon_0\epsilon_r}(x^2 - 2x_0 x) + C_2, \Rightarrow \\
x &= x_0, \phi(x_0) = \frac{E_c}{q} \Rightarrow C_2 = \frac{E_c}{q} - \frac{qN_d x_0^2}{2\epsilon_0\epsilon_r}, \frac{E_c}{q} = \text{ potential outside depletion layer} \\
\phi(x) &= \frac{E_c}{q} - \frac{qN_d}{2\epsilon_0\epsilon_r}(x - x_0)^2 \Rightarrow x = 0, \phi(0) = \frac{E_c}{q} - \frac{qN_d x_0^2}{2\epsilon_0\epsilon_r} = \frac{E_c}{q} - \Phi_{bi} \qquad (4.12) \\
\Phi_{bi} &= \frac{qN_d x_0^2}{2\epsilon_0\epsilon_r} \Rightarrow \\
x_0 &= \left(\frac{2\epsilon_0\epsilon_r\Phi_{bi}}{qN_d}\right)^{\frac{1}{2}} \propto N_d^{\frac{1}{2}} \qquad (4.13)
\end{aligned}$$

Clearly the thickness of the depletion layer is dependent on the square root of the doping concentration and the dielectric constant.

Note: In figure (4.3) the potential curves refer to the *electron potentials*, the expressions for the bending of the potential of the lower boundary of the conduction band and the upper boundary of the valence band are:

$$\begin{aligned}
\phi_{el}(x) &= \frac{E_c}{q} + \frac{qN_d}{2\epsilon_0\epsilon_r}(x - x_0)^2, \quad \phi_{el}(0) = \frac{E_c}{q} + \frac{qN_d x_0^2}{2\epsilon_0\epsilon_r} = \frac{E_c}{q} + \Phi_{bi}, \quad \phi_{el}(x_0) = \frac{E_c}{q} \\
\phi_{el}(x) &= \frac{E_v}{q} + \frac{qN_d}{2\epsilon_0\epsilon_r}(x - x_0)^2, \quad \phi_{el}(0) = \frac{E_v}{q} + \frac{qN_d x_0^2}{2\epsilon_0\epsilon_r} = \frac{E_v}{q} + \Phi_{bi}, \quad \phi_{el}(x_0) = \frac{E_v}{q}
\end{aligned}$$

If we were to assume a built-in potential $\Phi_{bi} = (U_M - U_S)$ of ≈ 1 eV and an $\epsilon_r = 12$ for Si, we find the following values for x_0: 1 μm ($N_d = 10^{21}$ m^{-3}), 0.1 μm ($N_d = 10^{23}$ m^{-3}), 0.01 μm ($N_d = 10^{25}$ m^{-3}).

Note: This is the expression for the extent of the depletion layer for an unbiased junction, featuring the built-in potential of the junction. If we were to apply an external voltage V_{ex} to the junction, Φ_{bi} should be replaced by $(\Phi_{bi} + V_{ex})$, V_{ex} taken positive for reverse bias and negative for forward bias, giving rise to a voltage dependent $(\Phi_{bi} + V_{ex})^{\frac{1}{2}}$ depletion depth. Moreover, like a vacuum diode, a Schottky diode acts as a rectifier when applying an external voltage, see figure (4.2). Therefore, in first approximation, the junction can be considered as a voltage dependent capacitor in parallel with a non-linear resistor in series with the bulk resistance of the semiconductor. Using the expression for the capacitance of a parallel plate capacitor $C = A(\epsilon_0\epsilon_r)/d$ (A = area, d = plate distance) and substituting $d = x_0$, we find for the *depletion layer* (or *'transition'*), capacitance:

$$C_{dep} = A\left[\frac{\epsilon_0\epsilon_r qN_d}{2(\Phi_{bi} + V_{ex})}\right]^{\frac{1}{2}} \quad \text{also} \quad \frac{1}{C_{dep}^2} = \frac{2(\Phi_{bi} + V_{ex})}{\epsilon_0\epsilon_r qN_d A^2} \propto (\Phi_{bi} + V_{ex}) \qquad (4.14)$$

Figure (4.4) shows the linear relation between $1/C_{dep}^2$ and the bias voltage V_{ex} that can

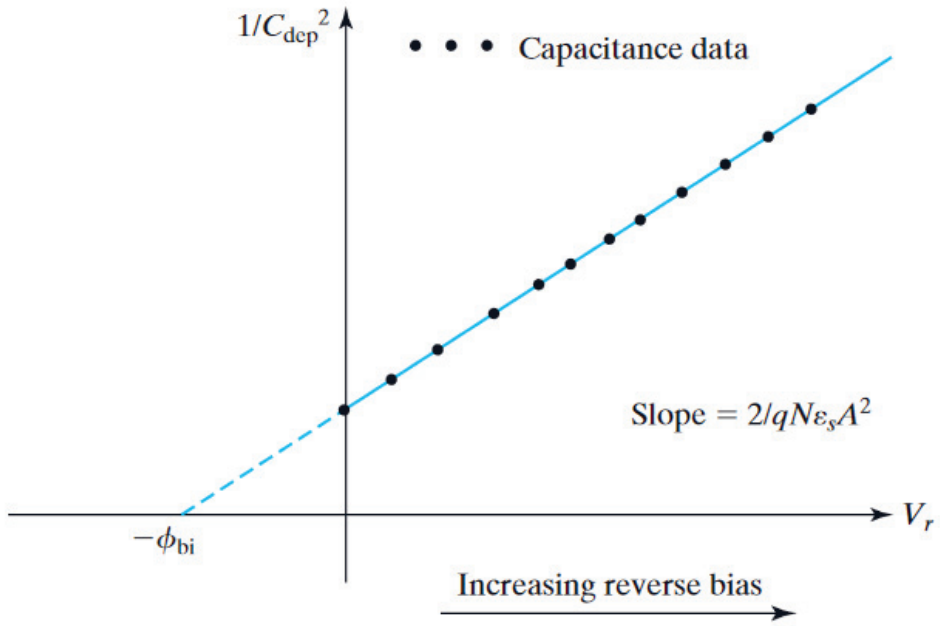


Figure 4.4: *Linear relation between the inverse square of the capacitance of the depletion layer (transition capacitance) C_{dep} and the junction bias voltage. The intersection with the abcissa yields the value of the the built-in potential Φ_{bi}, from the slope of the line one can derive the value of the doping concentration N_d of the n-type semiconductor.*

be experimentally determined by measuring the capacitance of the depletion layer at various values of the bias voltage (black dots). From this measurement we can extract the value of the built-in potential Φ_{bi} (intersection with the reverse-bias axis) and the value of the doping concentration N_d of the semiconductor layer if the area A of the junction contact is known. We can also compute the value of Φ_{bi} from equation (4.10) and hence calculate the value of the depletion layer capacitance C_{dep}.
Example: Find the capacitance of an $A = 10^{-9}$ m^2 *unbiased* Platinum/Silicon (Pt/Si) junction with the Si-layer doped to $N_d = 10^{22}$ m^{-3}. We have:

$$\begin{aligned}
\Phi_{B_n} &= \Phi_{bi} + \frac{kT}{q}\ln\left(\frac{g_c}{N_d}\right) \quad \text{with } q\Phi_{B_n} = 0.9 \text{ [eV] for Pt/Si} \quad \Rightarrow \\
0.9 \text{ [eV]} &= q\Phi_{bi} + 0.0258 \text{ [eV]} \ln\left(\frac{2.8\cdot 10^{25} \text{ [m}^{-3}\text{]}}{10^{22} \text{ [m}^{-3}\text{]}}\right) \quad \Rightarrow \quad q\Phi_{bi} = 0.69 \text{ [eV]} \\
\Phi_{bi} &= 0.69 \text{ [Volt]} \quad \text{for the above } \textit{unbiased} \text{ Pt/Si junction} \qquad (4.15) \\
C_{dep} &= A\left(\frac{\epsilon_0\epsilon_r q N_d}{2\Phi_{bi}}\right)^{\frac{1}{2}} \\
&= 10^{-9} \text{ [m}^2\text{]}\left(\frac{8.85\cdot 10^{-12} \text{ [F/m]}\cdot 12\cdot 1.6\cdot 10^{-19} \text{ [Coulomb]}\cdot 10^{22} \text{ [m}^{-3}\text{]}}{2\cdot 0.69 \text{ [Volt]}}\right)^{\frac{1}{2}} \quad \Rightarrow \\
&= 0.35\cdot 10^{-12} \text{ [Farad]} = 0.35 \text{ [pF]} \qquad (4.16)
\end{aligned}$$

As we shall see later, when describing a pn-junction, the Schottky diode, thanks to a narrower depletion region, has a substantially lower forward voltage drop: 0.3-0.4 Volt as compared to 0.7 Volt for a Si pn-junction. This can be beneficial in applications where power savings are an absolute must, such as battery driven and solar cell applications. Moreover, since the Schottky junction is unipolar (i.e. only majority charge carriers) it is apt for high speed switching applications where the diode needs to turn on and off quickly between forward and reverse bias and is not being slowed by the settling time of minority charge carriers. This makes Schottky diodes very useful for high frequency RF-applications and in high-speed switching power circuits with a typical recovery time between about 100 picoseconds and 1 nanosecond. For high frequency applications it is essential that the junction capacitance be kept as small as possible to avoid that the capacitive impedance $(1/j\omega C)$ causes so much leakage over the junction that the I-V characteristics are seriously impaired. Intrinsically, since the *Schottky junction is solely a majority charge carrier device*, it does not have any *diffusion* (or *'storage'*) capacitance C_D that can largely dominate the total capacitance $(C_{dep} + C_D)$ of a pn-junction in *forward* bias. The diffusion capacitance originates from the capacitive effect of the injection and storage of minority charge carriers into the body of the semiconductor from the opposite side of the junction *when in forward bias*. We shall discuss this in some detail in the next paragraph when treating the pn-junction. In that sense the Schottky contact is intrinsically very fast. Moreover, as shown in the example above, both minimizing the junction area (A) and selecting the minimum allowable doping concentration (N_d) are the device parameters that need to be optimized in unison to maximally reduce the capacitance of the depletion layer for good ultra high frequency (several GHz) performance.
The down side of a relatively narrow depletion region is that Schottky diodes are prone to high leakage currents when reverse biased and cannot withstand high reverse voltages in comparison to a pn-junction diode.

4.4 Semiconductor-semiconductor contact: the pn-junction

For the pn-junction we assume that an abrupt transition exists between the n-type and the p-type semiconductor material. The requirement for thermodynamical equilibrium implies that the Fermi level at both sides of the transition has to be equalized. As a consequence we have a concentration gradient for electrons relative to the p-type material and similarly for holes relative to the n-type material. This will activate a free charge transport across the contact barrier resulting in a transition region around the contact that contains practically no free charge carriers. This transition region is characterized by two space charge regions:

- A positive space charge owing to the presence of the ionized donor atoms locked in the lattice of the n-type material.

- A negative space charge owing to the presence of the ionized acceptor ions locked in the lattice of the p-type material.

Since this transition region is almost void of any free charge carriers, it is commonly called the *depletion layer*. The distribution of locked and free charges in the junction is shown in figure (4.5). In this simple model it is assumed that we also have abrupt transitions between the depletion layer and the bulk of the semiconductor that contains free charges. The space charge in the depletion region causes a built-in potential Φ_{bi} in an unbiased pn-junction the magnitude of which is calculated in what follows.

Suppose we have equilibrium concentrations n_1, p_1 in the n-type material and n_2, p_2 in the p-type material. Because of minority carrier suppression we also have $n_1 \gg n_2$ and $p_2 \gg p_1$. Because of the concentration gradients across the contact we shall have *diffusion current densities* $\vec{j}_{dn}$ for electrons and $\vec{j}_{dp}$ for holes that add up to a total diffusion current density across the contact barrier of $\vec{j}_d = \vec{j}_{dn} + \vec{j}_{dp}$. At equilibrium, when no net current can flow, the diffusion current density should be compensated for by a *field current density* $\vec{j}_c = \vec{j}_{cn} + \vec{j}_{cp}$ that originates from thermally generated electron-hole pairs that drift with velocity $\vec{v}_{drift} = \mu\vec{E}$ in the space charge electric field $\vec{E}$. It is evident that the diffusion current density will be proportional to the magnitude of the

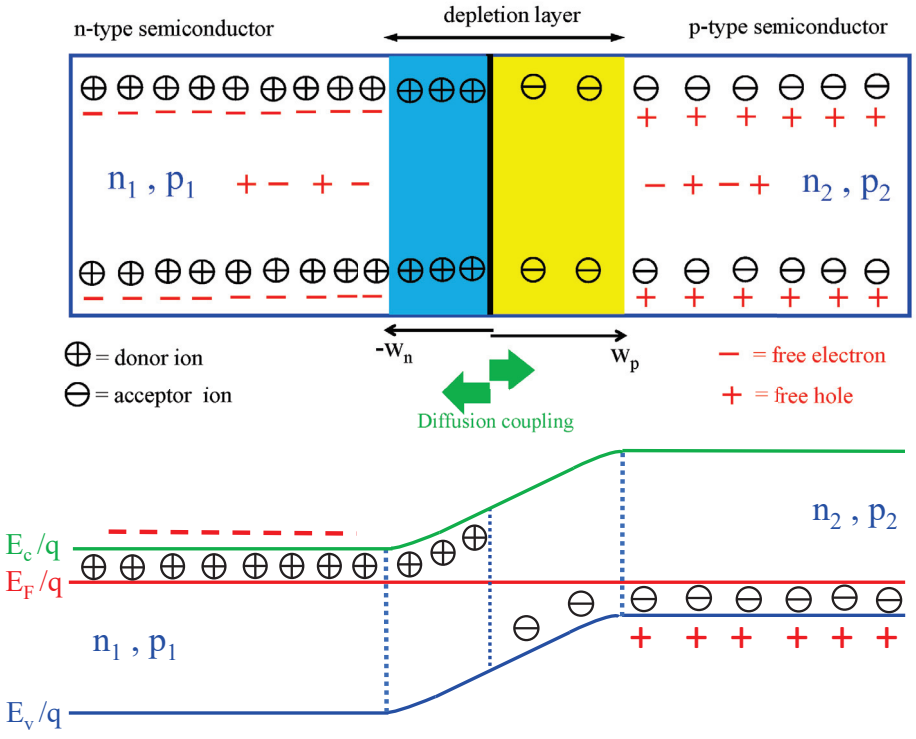


Figure 4.5: *Upper picture: The creation of a depletion layer in a pn-contact. The blue and yellow regions represent a positive and a negative space charge respectively, made up of positive and negative impurity ions that are locked in the crystal lattice. The free charge carriers (− for electrons and + for holes) are also indicated. Lower picture: band diagram of the pn-junction after equalization of the Fermi levels. The bending of the conduction band and the valence band as a result of the space charges are indicated (beware: electron potentials).*

concentration gradient (= change in concentration per unit length). Let's now consider, for simplicity, a situation whereby the concentration only varies in the x-direction and charge carrier concentrations in planes perpendicular to the x-axis do not occur. In that case the electron current density as a consequence of diffusion is proportional to dn/dx and the hole current density is proportional to dp/dx. The four relevant current densities are shown in figure (4.6) for three pn-bias situations (no external bias, reverse bias and forward bias). With no external bias and full compensation we have the following relations for the diffusion and field current densities:

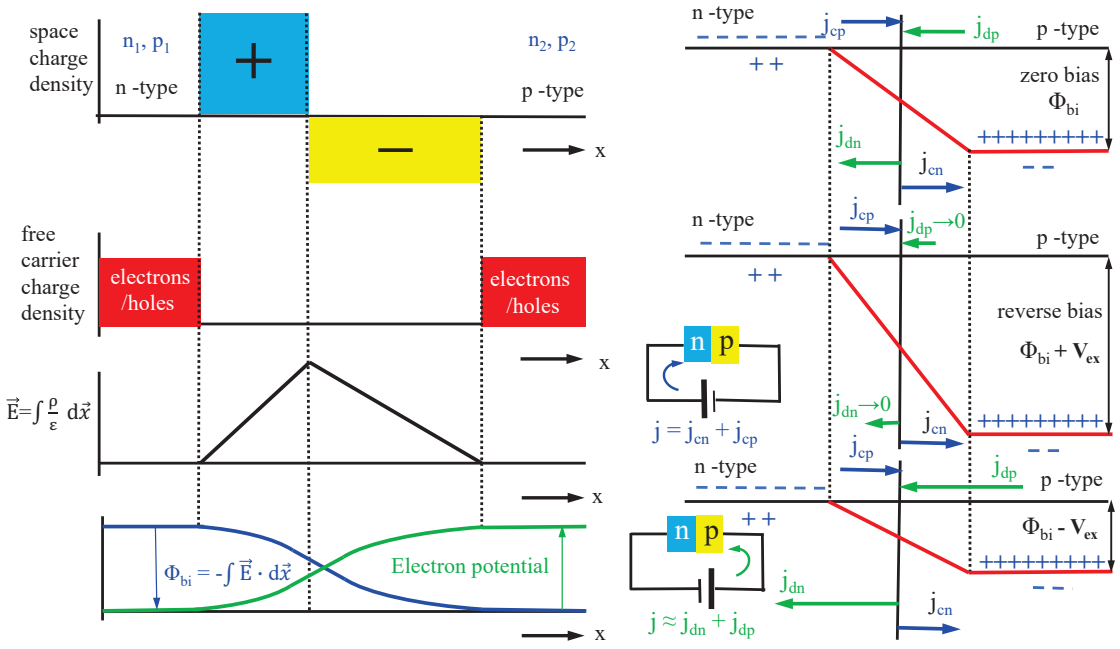


Figure 4.6: *Left: Space charge density, free charge carrier density, field strength E and potential Φ, i.e. electric (blue) and electron (green) potentials, are drawn-in in an idealized pn-junction (abrupt transitions). Right: Currents through a pn-junction. Upper picture: the equilibrium situation, $V_{ex} = 0$, no net current flowing. Middle picture: reverse bias $\Phi_{bi} + V_{ex}$, the diffusion current goes to practically zero. Lower picture: forward bias $\Phi_{bi} - V_{ex}$, holes from the p-region and electrons from the n-region are injected into the body of the semiconductor from the opposite side of the junction and the current through the junction is fully dominated by diffusion.*

$$j_n = j_{cn} + j_{dn} = 0 \quad \Rightarrow qn\mu_n E(x) = -qD_n \frac{dn}{dx} \Rightarrow \quad \frac{dn}{dx} = n\frac{\mu_n}{D_n}\frac{\phi(x)}{dx} \tag{4.17}$$

$$j_p = j_{cp} + j_{dp} = 0 \quad \Rightarrow qp\mu_p E(x) = qD_p \frac{dp}{dx} \quad \Rightarrow \quad \frac{dp}{dx} = -p\frac{\mu_p}{D_p}\frac{\phi(x)}{dx} \tag{4.18}$$

with $\phi(x)$ the electric potential as a function of position x across the junction. Applying Boltzmann's distribution function, we can also express the electron and hole densities as a function of the electric potential $\phi(x)$ and compare these with expressions (4.17)

and (4.18):

$$n = n_0\, e^{q\phi(x)/kT} \quad \Rightarrow \frac{dn}{dx} = n\frac{q}{kT}\frac{\phi(x)}{dx} \quad \Rightarrow \frac{\mu_n}{D_n} = \frac{q}{kT} \tag{4.19}$$

$$p = p_0\, e^{-q\phi(x)/kT} \quad \Rightarrow \frac{dp}{dx} = -p\frac{q}{kT}\frac{\phi(x)}{dx} \quad \Rightarrow \frac{\mu_p}{D_p} = \frac{q}{kT} \tag{4.20}$$

which resulted in the so-called Einstein relation between mobility, diffusion coefficient and temperature. To find the value of the built-in bias Φ_{bi} we can apply the Boltzmann relations of expressions (4.19) and (4.20) to compute the ratios n_2/n_1 and p_2/p_1 in the equilibrium situation:

$$\frac{n_2}{n_1} = e^{-q\Phi_{bi}/kT} \quad \text{or} \quad \frac{p_2}{p_1} = e^{q\Phi_{bi}/kT}, \quad \text{and also} \quad \frac{n_2 p_2}{n_1 p_1} = \frac{n_{i1}^2}{n_{i2}^2} = 1, \tag{4.21}$$

both expressions are equivalent since the intrinsic concentrations n_{i1} and n_{i2} should be equal when the temperatures at each side of the junction contact are the same. If the doping concentrations N_d and N_a are known and if we assume that at $T \approx 300K$ practically all donor and acceptor impurities are ionized, we have $n_1 = N_d$ and $p_2 = N_a$. Using the ratios of (4.21) we get:

$$\frac{q\Phi_{bi}}{kT} = \ln\frac{p_2}{p_1} = \ln\frac{p_2 n_1}{n_i^2} = \frac{N_d N_a}{n_i^2} \quad \Rightarrow \quad \Phi_{bi} = \frac{kT}{q}\ln\frac{N_d N_a}{n_i^2} \tag{4.22}$$

For Ge, assuming doping concentrations $N_d = 2\cdot 10^{22}$ m^{-3} and $N_a = 10^{22}$ m^{-3} one finds $\Phi_{bi} = 0.34$ Volt, for Si $\Phi_{bi} = 0.73$ Volt.
So let's now assess the situation where an external bias V_{ex} is applied to the junction. The 'Right' panel in figure (4.6) shows the potential $\phi(x)$ across the junction, if the applied V_{ex} is positive, the potential drop will be lowered to $\Phi_{bi} - V_{ex}$ (forward bias), if the applied V_{ex} is *reversed to negative*, Φ_{bi} will further drop to $\Phi_{bi} + V_{ex}$ (reverse bias). In the forward bias case we shall have a substantial injection of minority carriers, i.e. electrons injected in the p-type material and holes in the n-type material. As long as the amount of injected charge carriers remains small as compared to the concentrations of the majority charge carriers, the overall characteristics of the bulk material will not be impaired in terms of the validity of the majority/minority charge carrier model. We may anticipate that in that case the new concentrations can be found by substituting $(\Phi_{bi} - V_{ex})$ instead of Φ_{bi} in expression (4.21). Hence, for the new concentrations n_2^* and p_1^* we write:

$$n_2^* = n_1\, e^{-q(\Phi_{bi}-V_{ex})/kT} \quad \text{and} \quad p_1^* = p_2\, e^{-q(\Phi_{bi}-V_{ex})/kT} \tag{4.23}$$

The new concentration for the electron component can now be written as $n_2^* = n_2 + \Delta n_2$, i.e. as the equilibrium concentration plus a differential increment Δn_2:

$$\Delta n_2 = n_1\left(e^{-q\Phi_{bi}/kT}\right)\left(e^{qV_{ex}/kT} - 1\right) \tag{4.24}$$

The diffusion current density of electrons in the p-type material should be proportional to this extra concentration Δn_2 that was injected, hence:

$$j_n = C\left(e^{qV_{ex}/kT} - 1\right) \quad \text{with } C \text{ a constant that depends on the temperature} \tag{4.25}$$

From this derivation it is evident that j_n is an extra component in the diffusion current density j_{dn} that originates from the electron injection in the p-region. The total diffusion current density then amounts to:

$$j_{dn} = C\left(e^{qV_{ex}/kT} - 1\right) + j_{dn}(V_{ex} = 0) \quad \text{in which } j_{dn}(V_{ex} = 0) = -j_{cn} \tag{4.26}$$

If the reverse bias voltage is further enhanced, the diffusion current density becomes increasingly pinched, i.e. $j_{dn} \to 0$, from equation (4.26) with $V_{ex} \to -\infty$ we get $C = -j_{cn}$:

$$\begin{aligned} j_n &= -j_{cn}\left(e^{qV_{ex}/kT} - 1\right), \quad \text{and similarly for the injection of holes in the n-region:} \\ j_p &= -j_{cp}\left(e^{qV_{ex}/kT} - 1\right), \quad \text{so that we arrive at a total current density:} \\ j_{pn} &= -\left(j_{cn} + j_{cp}\right)\left(e^{qV_{ex}/kT} - 1\right) \end{aligned} \tag{4.27}$$

So with a strongly negative V_{ex}, the only remaining current density that flows is $(j_{cn} + j_{cp})$, the so-called *saturation* current or *reverse current*. We can assign this reverse current as $j_s = -(j_{cn} + j_{cp})$, thus:

$$j_{pn} = j_s\left(e^{qV_{ex}/kT} - 1\right) \tag{4.28}$$

Equation (4.28) shows the relation between the current in the outside circuitry and the external voltage applied to the pn-junction. It is clear from this formula that the pn-contact has rectifying properties and behaves like a diode: the pn-diode. The behavior is similar to that of a metal-semiconductor contact like the Schottky diode, but also exhibits significant differences like the magnitude of the reverse current, the breakdown voltage and the junction capacitance. The latter we shall treat in the following paragraph.

The junction capacitance comprises two components. One component is due to the narrowing and widening of the depletion layers as a function of junction voltage, the so-called *transition* capacitance, the most important parameters here are the actual width of the barrier and the junction material and its structure. The second component concerns the so-called *diffusion* or *storage* capacitance, that arises with forward bias when a large density of minority charge carriers is injected into the body of the semiconductor from the opposite side of the junction. This density is high near the junction, but trails off when moving away from the transition region.

Let's first assess the transition capacitance, for this we need to calculate the width of the space charge region that constitutes the depletion layer as a function of junction bias and impurity concentrations in the p- and n-regions. To handle this mathematically in a simple way, we shall assume that the junction is abrupt and that the depletion layers are completely depleted and that they. also, end abruptly. The assumption of complete depletion amounts to neglecting any left free carrier concentrations $n(x)$ and $p(x)$, each of which is much smaller than the impurity concentrations N_d and N_a which is entirely valid in practical n-doped or p-doped semiconductors. The space charge density, the associated electric field and electric potentials were also shown in figure (4.6). Let us first assume that we have as yet no external bias ($V_{ex} = 0$). Take the origin of the x-as at the position of the pn-contact plane, the extent of the depletion layer in the n-type

region x_n and the extent in the p-type region x_p. We can now derive expressions for x_n and x_p by solving the Poisson equation for the space charge regions in the n-type and p-type material:

$$\frac{d^2\phi(x)}{dx^2} = -\frac{\rho}{\epsilon_0\epsilon_r} \quad \text{with } \rho \text{ the space charge density} \tag{4.29}$$

For the depleted n-type region ($-x_n < x < 0$) containing positively charged donor ions with $\rho = \rho_+ = +qN_d$ we have:

$$\begin{aligned}
\frac{d^2\phi(x)}{dx^2} &= -\frac{qN_d}{\epsilon_0\epsilon_r}, \Rightarrow \text{integration} \Rightarrow \frac{d\phi(x)}{dx} = -\frac{qN_d x}{\epsilon_0\epsilon_r} + C_1 = -E(x), \\
x &= -x_n \Rightarrow E(-x_n) = 0 \Rightarrow C_1 = -\frac{qN_d x_n}{\epsilon_0\epsilon_r} \Rightarrow \\
E(x) &= \frac{qN_d}{\epsilon_0\epsilon_r}(x + x_n) \Rightarrow \phi(x) = -\int_+ E(x)dx \Rightarrow \\
\phi(x) &= -\frac{qN_d}{2\epsilon_0\epsilon_r}(x^2 + 2x_n x) + C_2, \Rightarrow x = -x_n,\ \phi(-x_n) = \frac{E_c}{q}, \Rightarrow \\
C_2 &= \frac{E_c}{q} - \frac{qN_d x_n^2}{2\epsilon_0\epsilon_r} \Rightarrow \phi(x) = \frac{E_c}{q} - \frac{qN_d}{2\epsilon_0\epsilon_r}(x + x_n)^2 \\
\phi(0) &= \frac{E_c}{q} - \frac{qN_d x_n^2}{2\epsilon_0\epsilon_r} = \frac{E_c}{q} - \Phi_n \Rightarrow \quad \Phi_n = \frac{qN_d x_n^2}{2\epsilon_0\epsilon_r}
\end{aligned} \tag{4.30}$$

where Φ_n represents the potential drop over the n-type part of the depletion region. Similarly, for the depleted p-type region ($0 < x < +x_p$) containing negatively charged acceptor ions with $\rho = \rho_- = -qN_a$ we can derive:

$$\begin{aligned}
\frac{d^2\phi(x)}{dx^2} &= -\frac{(-qN_a)}{\epsilon_0\epsilon_r}, \quad \Rightarrow \frac{qN_a}{\epsilon_0\epsilon_r} \Rightarrow \\
E(x) &= -\frac{qN_a}{\epsilon_0\epsilon_r}(x - x_p) \Rightarrow \\
\phi(x) &= \frac{qN_a}{2\epsilon_0\epsilon_r}(x - x_p)^2 + C_2, \Rightarrow x = 0,\ \phi(0) = \frac{E_c}{q} - \Phi_n \Rightarrow \\
C_2 &= \frac{E_c}{q} - \Phi_n - \frac{qN_a x_p^2}{2\epsilon_0\epsilon_r}, \Rightarrow \quad \Phi_p = \frac{qN_a x_p^2}{2\epsilon_0\epsilon_r} \\
\phi(x) &= \frac{E_c}{q} + \frac{qN_a}{2\epsilon_0\epsilon_r}(x - x_p)^2 - \Phi_{bi}, \Rightarrow \quad \Phi_{bi} = \Phi_n + \Phi_p, \text{ and} \\
\phi(x_p) &= \frac{E_c}{q} - \Phi_{bi}, \text{ with } \Phi_{bi} = \frac{q}{2\epsilon_0\epsilon_r}\left(N_d x_n^2 + N_a x_p^2\right),
\end{aligned} \tag{4.31}$$

with Φ_{bi} representing the built-in potential of the *unbiased* pn-junction. *Note*: In figure (4.5) the *electron* potentials are shown that reflect the bending of the conduction and valence band potential levels. For the depletion intervals ($-x_n < x < 0$) and ($0 < x <$

$+x_p$) these are given, respectively, by:

$$\phi_{el}(x) = \frac{E_c}{q} + \frac{qN_d}{2\epsilon_0\epsilon_r}(x+x_n)^2, \quad \phi_{el}(x) = \frac{E_c}{q} + \Phi_{bi} - \frac{qN_a}{2\epsilon_0\epsilon_r}(x-x_p)^2 \tag{4.32}$$

$$\phi_{el}(x) = \frac{E_v}{q} + \frac{qN_d}{2\epsilon_0\epsilon_r}(x+x_n)^2, \quad \phi_{el}(x) = \frac{E_v}{q} + \Phi_{bi} - \frac{qN_a}{2\epsilon_0\epsilon_r}(x-x_p)^2 \tag{4.33}$$

In addition we have the requirement of charge neutrality:

$$x_n qN_d + x_p qN_a = 0 \Rightarrow x_n qN_d = -x_p qN_a, \quad \frac{|x_n|}{|x_p|} = \frac{N_a}{N_d}, \quad \left(\text{also: } \frac{\Phi_n}{\Phi_p} = \frac{N_a}{N_d}\right) \tag{4.34}$$

From the equalities (4.31) and (4.34) we can derive the following expressions for x_n and x_p:

$$x_p = \left[\frac{2\epsilon_0\epsilon_r\Phi_{bi}}{qN_a(1+\frac{N_a}{N_d})}\right]^{\frac{1}{2}}, \quad x_n = \left[\frac{2\epsilon_0\epsilon_r\Phi_{bi}}{qN_d(1+\frac{N_d}{N_a})}\right]^{\frac{1}{2}} \tag{4.35}$$

$$x_{bi} = x_n + x_p = \left[\frac{2\epsilon_0\epsilon_r\Phi_{bi}}{q}\left(\frac{1}{N_d}+\frac{1}{N_a}\right)\right]^{\frac{1}{2}} \tag{4.36}$$

Hence, the width of the space charge region, at a certain doping concentration, varies with the square root of built-in contact potential Φ_{bi}. If we apply an external bias V_{ex}, the total contact potential will change to $\Phi_{pn} = \Phi_{bi} + V_{ex}$, V_{ex} being negative with forward bias and positive with reverse bias. In that case the width of the depletion layer will vary with $\Phi_{pn}^{1/2}$, i.e. substitute Φ_{pn} for Φ_{bi} in the above expressions.
For a symmetric space charge region in a *biased* pn-junction featuring $N_d = N_a = N$ we have for the electric field $E(x)$, the electric potential $\phi(x)$ and the *total width* w of the depletion layer the following relations:

$$\begin{aligned} E(x) &= \frac{qNw}{2\epsilon_0\epsilon_r}\Lambda\left(\frac{2x}{w}\right) \quad (\Lambda = \text{triangle function}) \\ \phi\,(-w/2 \le x \le 0) &= -\frac{qN}{2\epsilon_0\epsilon_r}(x+w/2)^2, \; \phi\,(0 \le x \le +w/2) = \frac{qN}{2\epsilon_0\epsilon_r}(x-w/2)^2 - \Phi_{pn} \\ w &= \left(\frac{4\epsilon_0\epsilon_r\Phi_{pn}}{qN}\right)^{\frac{1}{2}} \Rightarrow \quad \propto N^{-\frac{1}{2}}\Phi_{pn}^{\frac{1}{2}} \end{aligned} \tag{4.37}$$

If the conductance is, for example, high at the p-side of the junction as compared to the n-side, the depletion layer will be predominantly present at the low conductance n-side of the pn-junction, thus we have then $N_a \gg N_d$:

$$w \approx x_n = \left[\frac{2\epsilon_0\epsilon_r\Phi_{pn}}{qN_d}\right]^{\frac{1}{2}} \tag{4.38}$$

This expression is exactly the same as formula (4.13) for the width of the depletion layer (x_0) in a Schottky contact. This is of course to be expected, since the high conductance, heavily doped, semiconductor contact in the pn-junction acts as the high conductivity

metal contact present in the Schottky junction.
With fully ionized donor impurities at the n-side, we have a conductance $\sigma_n = \mu_n q N_d$ and hence a barrier width:

$$x_{pn} = \left(\frac{2\mu_n \epsilon_0 \epsilon_r \Phi_{pn}}{\sigma_n} \right)^{\frac{1}{2}} \tag{4.39}$$

We might, for example, consider the width of the barrier for a Silicon pn-junction with a conductivity $\sigma_n = 200$ mhos m^{-1}. For the width of the depletion layer we then find $w_n = (2 \cdot 0.14 \cdot 8.854 \cdot 10^{-12} \cdot 12 \cdot \Phi_{pn}/200)^{1/2} \approx 0.4\ \Phi_{pn}^{1/2}\ \mu$m.
There is a change of charge associated with a change in depletion layer width, and since the latter is voltage dependent we can also associate a capacitance with the depletion layer. In figure (4.6) we show a total positive charge per unit area $Q_+ = +w_n q N_d$ in the n-type Si and a total negative charge per unit area $Q_- = -w_p q N_a$ in the p-type Si, because of charge neutrality $Q_+ + Q_- = 0 \rightarrow Q_+ = -Q_-$, so we have charges Q with opposite polarity stored on either side of the junction. These charges vary with the applied external bias V_{ex}. If we consider now specifically the depletion region in the n-type Si, we can define an associated transition capacitance as:

$$C_T = \frac{dQ_+}{dV_{ex}} = \frac{dQ_+}{d\Phi_{pn}} = \frac{dQ_+}{dw_n} \cdot \frac{dw_n}{d\Phi_{pn}} \tag{4.40}$$

Differentiating $Q_+ = Q$, the charge stored at either side of the junction, we get:

$$\begin{aligned} \frac{dQ}{dw_n} &= qN_d, \quad \text{from differentiating (4.35):} \quad \frac{dw_n}{d\Phi_{pn}} = \frac{1}{2}\left[\frac{2\epsilon_0\epsilon_r}{qN_d(1+\frac{N_d}{N_a})} \right]^{\frac{1}{2}} \Phi_{pn}^{-\frac{1}{2}} \Rightarrow \\ C_T &= \left(\frac{q\epsilon_0\epsilon_r N_a N_d}{2} \right)^{\frac{1}{2}} (N_d + N_a)^{-\frac{1}{2}} \Phi_{pn}^{-\frac{1}{2}} \end{aligned} \tag{4.41}$$

The transition capacitance C_T is inversely proportional to the square root of the junction voltage $\Phi_{pn} = \Phi_{bi} + V_{ex}$ and since V_{ex} is positive for reverse bias, C_T will decrease if the reverse bias on the junction is raised. The maximum transition capacitance will therefore occur with forward bias (V_{ex} being negative), however V_{ex} will always remain smaller than Φ_{bi} so that the transition capacitance cannot become excessively large. As an example let's compute the value of C_T for a silicon junction with Φ_{bi} =0.7 Volt and a forward bias of 0.5 Volt. Let's furthermore take $N_d = 2 \cdot 10^{22}$ m^{-3} and $N_a = 10^{22}$ m^{-3}, we then find:

$$\begin{aligned} C_T &= \left(\frac{1.6 \cdot 10^{-19}[\text{Coulomb}] \cdot 8.854 \cdot 10^{-12}[\text{F/m}] \cdot 12 \cdot 2 \cdot 10^{22}[\text{m}^{-3}] \cdot 10^{22}[\text{m}^{-3}]}{2} \right)^{\frac{1}{2}} \\ &\cdot \left(10^{22}[\text{m}^{-3}] + 2 \cdot 10^{22}[\text{m}^{-3}] \right)^{-\frac{1}{2}} \cdot (0.2[\text{Volt}])^{-\frac{1}{2}} \qquad [\text{Farad m}^{-2}] \\ C_T &= 5.3 \cdot 10^{-4}\,[\text{F m}^{-2}], \text{ with a junction cross section of } 10^{-8}\ [\text{m}^2] \Rightarrow C_T = 5.3\ [\text{pF}] \end{aligned}$$

However, with forward bias the diffusion capacitance normally dominates the junction capacitance and potentially impairs the rectifying characteristics of the pn-junction at high frequencies due to the low impedance ($1/j\omega C$) capacitive bypass. Hence we shall now briefly analyze the diffusion capacitance of a pn-junction.

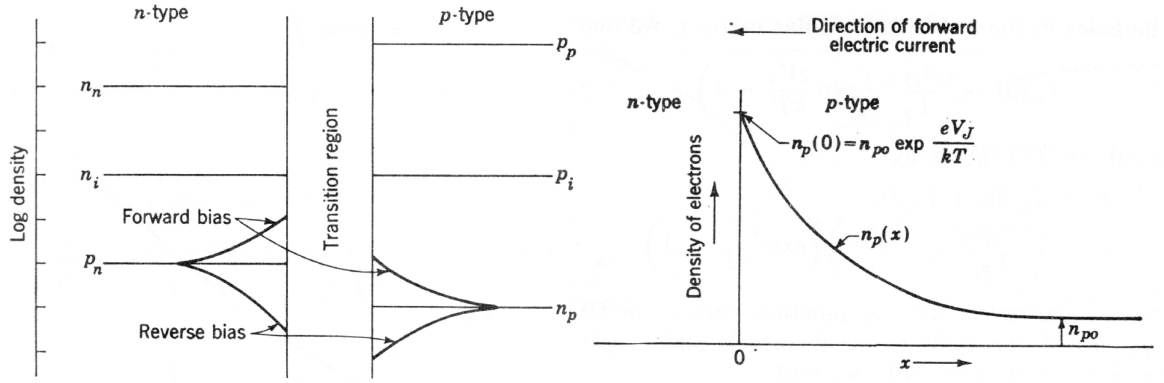


Figure 4.7: *Left: The carrier densities at two sides of the transition region are shown. The enrichment or depletion of the carriers near the junction is shown for forward and reverse bias respectively. Right: The carrier concentration of injected carriers and the boundary conditions needed to solve the diffusion problem and calculate the storage capacitance.*

With forward bias V_{ex}, a large density of minority charge carriers is injected into the body of the semiconductor from the opposite side of the junction. The density of these minority carriers will be high near the junction and will trail off as one moves away from the transition region. The distribution of the carrier concentration can be found by solving the diffusion equation. The situation is depicted in figure (4.7). The equilibrium minority density in the p-type region is indicated by n_{p0}, the distribution of the trailing excess minority carriers is marked by $n_p(x)$ that follows from the solution of a one-dimensional diffusion equation in coordinate x:

$$n_p(x) = n_{p0}\left(e^{qV_{ex}/kT} - 1\right)\, e^{-x/L_n} + n_{p0} \quad \text{with } L_n = (D_n\tau_n)^{\frac{1}{2}} \tag{4.42}$$

D_n and τ_n are the electron diffusion coefficient and life time in p-type silicon respectively, the boundary conditions are $n_p(0) = n_{p0}\, e^{qV_{ex}/kT}$ at $x = 0$ and $n_p(x) = n_{p0}$ at $x = \infty$. The electron excess distribution is then given by:

$$n'_p(x) = n_{p0}\left(\, e^{qV_{ex}/kT} - 1\right)\, e^{-x/L_n} \tag{4.43}$$

It is the total charge in this excess tail of minority carriers (to be neutralized by majority carriers from the external circuit), which contributes to the diffusion or storage

capacitance. Integration of (4.43) yields:

$$Q_- = -q\int_0^\infty n_p'(x)dx = -qn_{p0}\left(e^{qV_{ex}/kT}-1\right)\int_0^\infty e^{-x/L_n}dx = -qn_{p0}L_n\left(e^{qV_{ex}/kT}-1\right) \tag{4.44}$$

If the external voltage V_{ex} is varied, the differential capacitance due to the change in stored electron charge is given by:

$$\begin{aligned} C_- &= \frac{-dQ_-}{dV_{ex}} = \frac{q^2 n_{p0}L_n}{kT}\,e^{qV_{ex}/kT}, \quad \text{adding to this the change in stored hole charge:} \\ C_D &= C_- + C_+ = \frac{q^2}{kT}\,e^{qV_{ex}/kT}\left(n_{p0}L_n + p_{n0}L_p\right) \end{aligned} \tag{4.45}$$

It is clear that the total diffusion capacitance C_D exponentially depends on the forward bias V_{ex} (in fact linearly depends on the forward diode current density j_{pn}) and can potentially become much larger than the transition capacitance C_T. This notion is made more explicit in the following practical example.
Consider a Si pn-junction with a built-in potential $\Phi_{bi} = 0.73$ [Volt], cross sectional area A$= 10^{-8}$ [m^2] with a doping concentration $N_d = 2\cdot 10^{22}$ [m^{-3}] and $N_a = 10^{22}$ [m^{-3}] at an ambient temperature T =300 K (thermal potential $kT/q = 25.8\cdot 10^{-3}$ Volt). Furthermore we have the following values for the diffusion constants: $D_n = 2.5\cdot 10^{-3}$ [m^{2}sec^{-1}], $D_p = 10^{-3}$ [m^{2}sec^{-1}] and life times: $\tau_n = 5\cdot 10^{-7}$ [sec], $\tau_p = 10^{-7}$ [sec].
The equilibrium concentrations of the minority charge carriers and the effective diffusion lengths can be now be obtained from:

$$\begin{aligned} n_{p0} &= n_{n0}\,e^{-q\Phi_{bi}/kT} \quad \text{with } n_{n0} = N_d \quad \Rightarrow n_{p0} = 10^{10}\ [\text{m}^{-3}]\left(=\frac{n_i^2}{N_a}\right) \\ p_{n0} &= p_{p0}\,e^{-q\Phi_{bi}/kT} \quad \text{with } p_{p0} = N_a \quad \Rightarrow p_{n0} = 5\cdot 10^{9}\ [\text{m}^{-3}]\left(=\frac{n_i^2}{N_d}\right) \\ L_n &= \sqrt{D_n\tau_n} = 3.5\cdot 10^{-5}\,[\text{m}] \\ L_p &= \sqrt{D_p\tau_p} = 10^{-5}\,[\text{m}] \end{aligned}$$

Substituting the above numerical values in the expression for C_D, we get:

$$\begin{aligned} C_D &= 2.48\cdot 10^{-20}e^{V_{ex}/(25.8\cdot 10^{-3})}[\text{Farad}], \quad \text{for three incremental values of } V_{ex} \Rightarrow \\ C_D &\approx 0.43\ [\text{pF}], \quad \text{forward bias } V_{ex} = 0.43\ [\text{Volt}] \\ C_D &\approx 21\ [\text{pF}], \quad \text{forward bias } V_{ex} = 0.53\ [\text{Volt}] \\ C_D &\approx 1.0\ [\text{nF}], \quad \text{forward bias } V_{ex} = 0.63\ [\text{Volt}] \end{aligned}$$

The exponential rise of C_D with V_{ex} shows that with a substantial forward current flow through the pn-diode, the diffusion capacitance can become orders of magnitude larger than the depletion capacitance C_T. In that case C_D actually completely dominates the high frequency response of the diode (i.e. tending to an ohmic contact), whereas this effect is completely absent in a Schottky diode.

Chapter 5

Sensor physics: noise sources

5.1 Physical foundations of noise sources

The most important noise sources in electronics comprise thermal noise and shot noise and we shall treat the physical foundations of these noise sources in some detail. After this we shall only briefly touch on a few other sources of noise without any elaboration. The various noise source signals cannot be readily identified on the basis of their characteristic wave shapes. The reason for this lies in the fact that they are always observed behind a frequency filter with a limited bandwidth that smooths out the original, potentially characteristic, features or details. This leads in all cases to the same statistical treatment with predominantly Gaussian probability density distributions.

5.2 Thermal (Nyquist, Johnson) noise

. Thermal noise includes the random fluctuation of an electric voltage over a resistor or resistive element in temperature equilibrium with its environment. The noise voltage originates from the thermal agitation of the electrons that move in thermal equilibrium with the vibrating molecules in the crystal lattice of the resistor material. Every randomly moving electron causes a small current in the resistor that generates a small voltage over the resistor terminals. The superposition of all these small random voltages constitutes a resultant thermal noise voltage $V_n(t)$ over the resistor. This thermal noise voltage is often referred to as Nyquist or Johnson noise, named after Harry Nyquist and John Johnson who pioneered the research of noise in electronics at Bell Labs in 1926.
In deriving the available noise power from such a thermal source, Nyquist developed an elegant explanation based on fundamental physics that we shall follow here.
Consider a loss-less electrical transmission line connected at one end to a resistor R (green) and at the other end to another resistor R_c, where $R_c = Z_0 = \sqrt{L/C} = R$ (red) with $Z_0 = R_c$ the characteristic impedance of the transmission line. We now have a closed circuit containing two equal resistors connected by a matched transmission line so that no reflection of electric waves traveling along the transmission line will occur, i.e. radiation arriving at the end of the line should be completely absorbed. In fact this set up may be regarded as a one-dimensional example of black body radiation and is depicted in figure 5.1. At a finite temperature T the green resistor, in equilibrium with

the surrounding thermal radiation field, generates a noise voltage $V_n(t)$ which will propagate down the transmission line. Arriving at the red resistor this electric wave should be fully absorbed, i.e. all frequencies in the propagating wave fields should have nodes at both ends of the transmission line implying an integer number of half-wavelengths over the line L. These permitted standing wave modes in the line have wavelengths $\lambda = 2L/n$, corresponding to frequencies $\nu = (v/2L)n$, where $n = 1, 2, 3, ...$ and v represents the wave phase velocity in the transmission line. The separation of subsequent modes in frequency equals $v/2L$ and the number of standing wave modes in an interval between ν and $\nu + d\nu$ amounts to:

$$m d\nu = \frac{2L}{v} d\nu \tag{5.1}$$

Thermal wave photons are bosons, so the energy distribution obeys Bose-Einstein statistics. We shall derive this statistical distribution in a next paragraph (see (5.26)) but utilize the result already here. The *mean* thermal energy $\overline{\epsilon}(\nu)$ contained in each electromagnetic standing wave mode (also called photon state) in the transmission line equals (Planck distribution):

$$\overline{\epsilon}(\nu) = \frac{h\nu}{e^{h\nu/kT} - 1} \tag{5.2}$$

with $h = 6.62\text{x}10^{-34}$ Joule.sec (Planck's constant), $k = 1.38\text{x}10^{-23}$ Joule/^{0}K (Boltzman's constant) and T (Kelvin) the temperature of the resistors R. The electromagnetic energy, integrated over frequency interval $d\nu$, amounts then to $m\overline{\epsilon}(\nu)d\nu$, half of this energy is generated by the green resistor R and propagates towards its equivalent at the

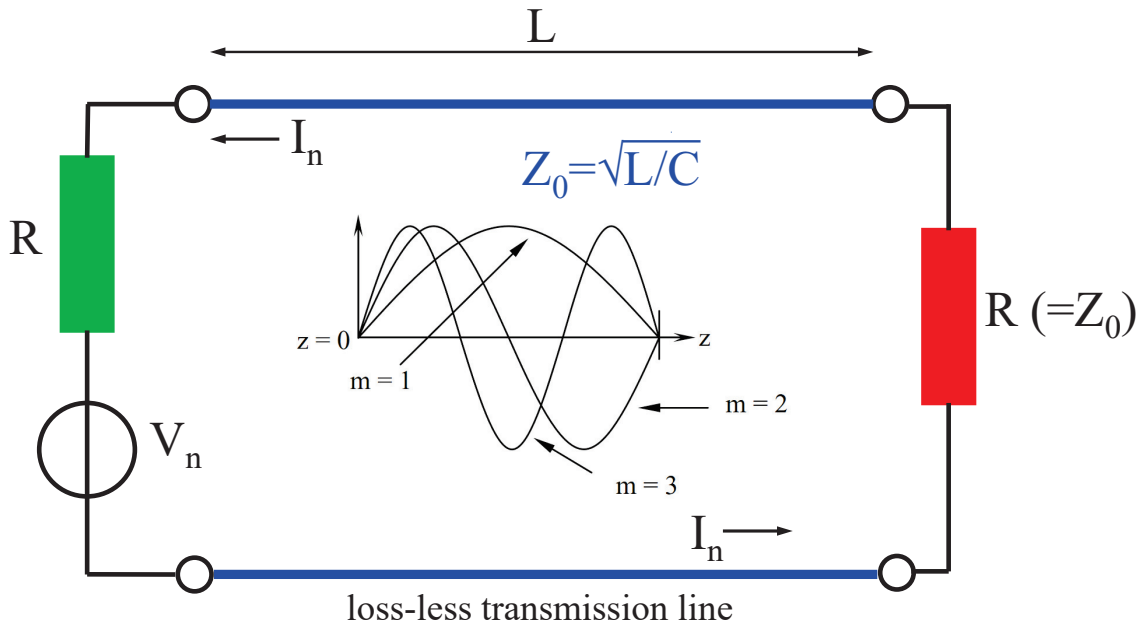


Figure 5.1: *Closed circuit comprising two equal resistors connected by a loss-less matched transmission line. The insert shows the first three standing wave modes.*

right, obviously the other half propagates from right to left. Since the propagation time Δt over the line from the generating resistor to the absorbing resistor is L/v, the absorbed *mean* energy per unit time, i.e. the *mean* power, equals $\overline{\epsilon}(\nu)Ld\nu/(v\Delta t)$ or:

$$\overline{P}(\nu)d\nu \;=\; \frac{h\nu}{e^{h\nu/kT}-1}\,d\nu, \quad \text{with } \overline{P}(\nu) \text{ the power spectral density [Watt Hz}^{-1}] \qquad (5.3)$$

This power, integrated over the relevant frequency bandwidth $\Delta\nu$, is simply the ohmic heating generated by the noise voltage source associated with the green resistor in figure 5.1. As explained earlier, the relation between the voltage noise source $V_n(t)$ and the optimal available power at the absorbing red resistor is:

$$W_n \;=\; \frac{\overline{V_n^2}}{4R} \;=\; \int_{\Delta\nu} \overline{P}(\nu)d\nu \quad \rightarrow \overline{V_n^2} = 4R\int_{\Delta\nu} \overline{P}(\nu)d\nu \;=\; \int_{\Delta\nu} \overline{S}_V(\nu)d\nu \text{ with:} \qquad (5.4)$$

$$\overline{S}_V(\nu) \;=\; 4R\,\frac{h\nu}{e^{h\nu/kT}-1} \quad \text{the voltage power spectral density [Volt}^2\text{Hz}^{-1}] \qquad (5.5)$$

Figure 5.2 shows $S_V(\nu)$ as a function of the frequency ν.

In the low frequency regime we may approximate $e^{h\nu/kT}$ by $(h\nu/kT+1)$, from which

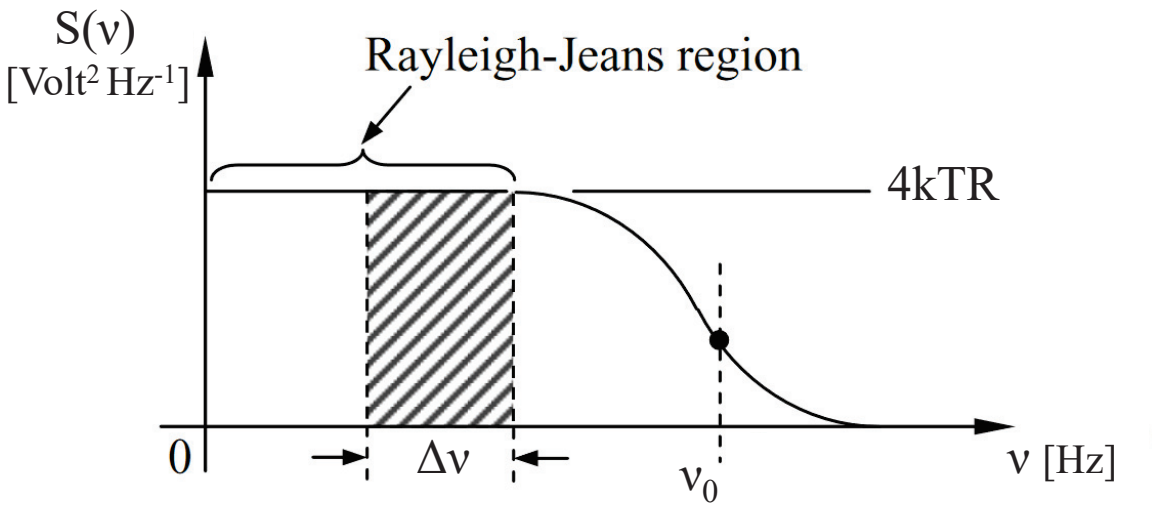


Figure 5.2: *The 'one sided' voltage power spectral density $S_V(\nu)$ as a function of frequency ν. If $h\nu \ll kT$, $S_V(\nu)$ reaches a plateau value of $4kTR$ [Volt^2 Hz^{-1}] independent of the frequency ν. This is the frequency range where the Rayleigh-Jeans approximation of the Planck distribution holds and is commonly referred to as the thermal limit.*

we get:

$$S_V(\nu) \;=\; 4kTR \qquad [\text{Volt}^2\ \text{Hz}^{-1}] \text{ for } \nu < \nu_0 \qquad (5.6)$$

This part of the power spectral density spectrum coincides with the Rayleigh-Jeans part of the Planck distribution for black body radiation and is valid up to $\nu_0 \approx kT/h$. At room temperature $T \approx 300$ K we have $\nu_0 \approx 6$ TeraHz (wavelength ≈ 50 μm), i.e. in the infrared waveband. For usage in electronics, approximation (5.6) is therefore fully

adequate. For frequencies $\nu > \nu_0$, $S(\nu)$ goes to zero. This is the region of the Planck distribution that coincides with the Wien branch that avoids the so-called ultraviolet catastrophe, quantum effects take over here as we shall see in the following paragraph. We can summarize the physical meaning of (5.6) as follows:

- Resistor noise exhibits a constant spectral density, independent of frequency, so-called 'white noise'.
- The magnitude of the noise power is proportional to the macroscopic quantities R and T. The noise power can be reduced by cooling the resistive element.
- If we consider a certain frequency bandwidth $\Delta\nu$, the total mean square noise voltage follows from:

$$\overline{V^2} \;=\; 4kTR \int_{\Delta\nu} S_V(\nu) d\nu \;=\; 4kTR\Delta\nu \qquad [\text{Volt}^2] \tag{5.7}$$

- The maximum available noise power follows from $\overline{P} = \overline{V^2}/4R = kT\Delta\nu$ [Watt], as explained in the previous paragraph. It is clear that the available noise power is independent of the actual value of the resistive element. Available noise powers play a major role in the assessment of noise disturbances in sensitive receiver systems, amplifiers and sensor elements.

The power spectral density for thermal noise can actually be obtained in various ways. In the previous section we have used the original method by Nyquist, based on the derivation of the black body Planck distribution. We shall present here one alternative method following an electronic approach that is based on coupling of the resistor into a tuned LCR circuit and on the application of the energy equipartition principle. In this approach the influence of quantum mechanical effects is neglected. The set-up is displayed in figure (5.3).

The differential equation that governs the relation between the voltage $V(t)$ and the current $I(t)$ can be written as:

$$V(t) \;=\; I(t)R + L\frac{dI(t)}{dt} + U(t), \qquad \text{with } U(t) \text{ the voltage over capacitor } C \tag{5.8}$$

From this we can find an energy equation (dropping now the time dependence in the notation) by multiplication with Idt, for the capacitor we have $Idt = dQ = CdU$, hence:

$$VIdt \;=\; I^2Rdt \;+\; LIdI + CUdU \;=\; I^2Rdt \;+\; d\left(\frac{L}{2}I^2 \;+\; \frac{C}{2}U^2\right) \tag{5.9}$$

Here we have $VIdt$ as the energy yield of the voltage source in dt (VI is the *momentaneous* source power P [Watt]) and I^2Rdt as the energy converted to heat in dt (I^2R is the *momentaneous* heating power [Watt]).
Consequently $(LI^2/2 + CU^2/2)$ should have the dimension of energy and $d(LI^2/2 + CU^2/2)$ comprises the increase in magnetic and electric field energy in dt (with $d(LI^2/2 + CU^2/2)/dt$ the *momentaneous* field power [Watt]).

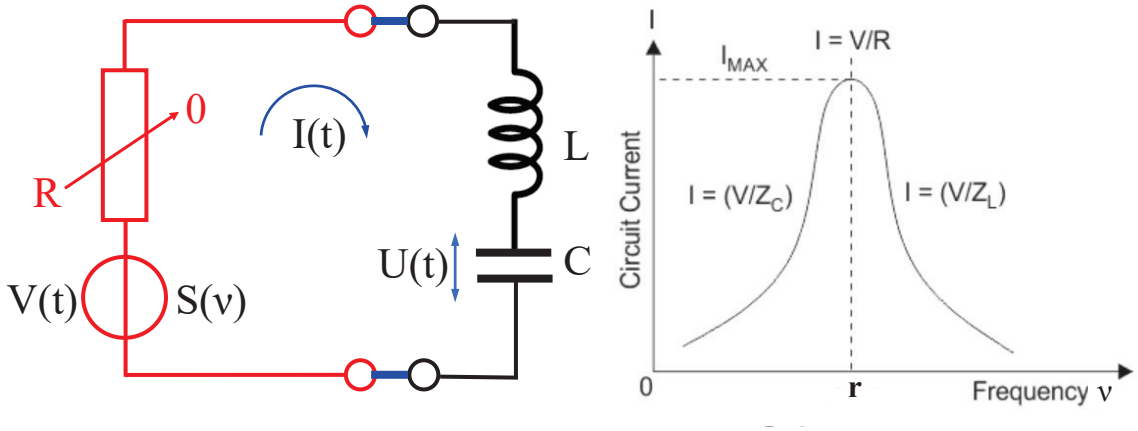


Figure 5.3: *A resistor noise source $V(t)$ feeding a series LCR circuit. The unknown voltage power spectral density distribution $S_V(\nu)$ can be assessed by chosing a small value for R that gives rise to a sharp peak (small frequency interval) at the circuit's resonance frequency.*

Since in this case $V(t)$ is a stationary noise voltage, we should use here the mean square value $\overline{V^2}$ and the mean square current $\overline{I^2}$. Hence the coil inductance L contains a *mean* magnetic energy $L\overline{I^2}/2$ and the capacitor C a *mean* electric field energy $C\overline{U^2}/2$. In the equilibrium situation we should have equipartition of energy between the various components, i.e.:

$$\frac{1}{2}kT \;=\; \frac{1}{2}L\overline{I^2} \;=\; \frac{1}{2}C\overline{U^2} \quad \text{with thermal energy } \tfrac{1}{2}kT \text{ for one degree of freedom} \tag{5.10}$$

From these equalities we apparently have $\overline{I^2} = kT/L$ independent of R and C and $\overline{U^2} = kT/C$ independent of R and L, e.g. for $L = 0$ and $R = \infty$.
Taking up the equality $\overline{I^2} = kT/L$, if we could succeed in relating $\overline{I^2}$ to $S_V(\nu)$ at a particular frequency ν then we would have arrived at an expression for $S_V(\nu)$. The LCR circuit has a complex impedance $\bar{Z}(j\omega) = R + j(\omega L - 1/\omega C)$ with $\omega = 2\pi\nu$. So we have for every radial frequency ω the following relation between $I(j\omega)$ and $V(j\omega)$:

$$|\bar{I}(j\omega)|^2 \;=\; \frac{|\bar{V}(j\omega)|^2}{|\bar{Z}(j\omega)|^2} \;=\; \frac{|\bar{V}(j\omega)|^2}{R^2 \,+\, (\omega L \,-\, 1/\omega C)^2} \tag{5.11}$$

The squared impedance $|\bar{Z}(j\omega)|^2$ has a minimum value at the resonance frequency ω_0 that follows from $\omega_0 L = 1/\omega_0 C$, hence $\omega_0 = 1/\sqrt{LC}$. At this frequency the squared current value $|\bar{I}(j\omega)|^2$ is at its maximum value. We can rewrite equation (5.11) as:

$$|\bar{I}(j\omega)|^2 \;=\; \frac{|\bar{V}(j\omega)|^2}{R^2\left[1 \,+\, \frac{L}{R^2C}\left(\frac{\omega}{\omega_0} \,-\, \frac{\omega_0}{\omega}\right)^2\right]} \tag{5.12}$$

To arrive at the mean square noise current we need to integrate over all frequencies, substitute $|\bar{V}(j\omega)|^2 = S_V(\nu)d\nu$ and $\omega = 2\pi\nu$:

$$\overline{I^2} \;=\; \int_0^\infty \frac{S_V(\nu)d\nu}{R^2\left[1 \;+\; \frac{L}{R^2C}\left(\frac{\nu}{\nu_0} \;-\; \frac{\nu_0}{\nu}\right)^2\right]} \tag{5.13}$$

If we take an increasingly smaller value for the resistor R the resonance peak gets narrower giving rise to a very sharp tuning of the LCR circuit on the resonance frequency ν_0. The contribution to the integral in (5.13) will then originate almost solely from a small region around ν_0, see the resonance peak displayed on the right side in figure 5.3. The sharply tuned LCR circuit in fact acts as a very selective spectrum analyzer, whereby $\overline{I^2}$ almost exclusively depends on the value of $S_V(\nu_0)$. It can be easily derived that if $\nu \to \nu_0$ the value of $[(\nu/\nu_0) - (\nu_0/\nu)] \to 2(\nu - \nu_0)/\nu_0$. Substitution in (5.13) and solving the integral yields:

$$\overline{I^2} \;=\; \frac{S_V(\nu_0)}{4\pi RL}\int_{-\infty}^{+\infty}\frac{dx}{1+x^2} \;\text{ with } x = \left[\tfrac{4\pi L}{R}(\nu - \nu_0\right] \to\; \overline{I^2} \;=\; \frac{S_V(\nu_0)}{4RL} \tag{5.14}$$

Note: The lower boundary of the integral over frequency shifts from $\nu = 0$ to $x_{low} = -(2\sqrt{L/C})/R$, resulting from the change of variables. If the value of R is lowered for a more selective tuning of the LCR-circuit to $S_V(\nu_0)$, we have in the limit $R \to 0$ that $x_{low} \to -\infty$. This leads to the value π for the definite integral.
From the equipartition requirement we already have $\overline{I^2} = kT/L$, so with the result from (5.14) we find:

$$S_V(\nu_0) \;=\; 4kTR \quad [\text{Volt}^2\text{Hz}^{-1}] \tag{5.15}$$

5.3 Inclusive treatment of thermal and quantum noise

The magnitude of fluctuations in an electromagnetic radiation field depends on whether the wave character or the quantum character (i.e. photons) dominates. By employing Bose-Einstein statistics, the magnitude of these intrinsic fluctuations can be computed for the specific case of a blackbody radiation field which is incoherent or "chaotic" with respect to time. Photons are *bosons*, which are not subject to the Pauli exclusion principle. Consequently many bosons may occupy the same quantum state.

5.3.1 Bose-Einstein statistics

The world is quantum-mechanical. A way to describe this is to note that particles (e.g. photons) in a unit volume of space ('a cubic meter') are distributed in momentum-space in boxes with a size proportional to h^3, where h is Planck's constant. At each energy, there is a finite number Z of boxes (where $Z \propto 4\pi p^2 dp$ with p the particle momentum). Consider n_i particles, each with energy ϵ_i, and call the number of boxes available at

that energy Z_i. Bosons can share a box. The number of ways $W(n_i)$ in which the n_i bosons can be distributed over the Z_i boxes is given by

$$W(n_i) = \frac{(n_i + Z_i - 1)!}{n_i!(Z_i - 1)!} \tag{5.16}$$

To digest this: the problem is equivalent to laying n_i particles and $Z_i - 1$ boundaries in a row. The total number of permutations is $(n_i + Z_i - 1)!$, but we have to account for the fact that the particles and boundaries may be interchanged. Similarly, we put n_j bosons in Z_j boxes, etc. The number of ways in which $N = \sum_{i=1}^{\infty} n_i$ bosons can be distributed over the boxes in momentum space thus is: $W = \prod_{i=1}^{\infty} W(n_i)$. The basic assumption of statistical physics is that the probability of a distribution is proportional to the number of ways in which this distribution can be obtained, i.e. to W. The physics behind this assumption is that collisions between the particles will continuously re-distribute the particles over the boxes. The most likely particle distribution is thus the one which can be reached in most different ways. We find this maximum in W by determining the maximum of $\ln W$, by setting its derivative to zero:

$$\ln W = \sum_{i=1}^{\infty} \ln W(n_i) \Rightarrow \Delta \ln W = \sum_{i=1}^{\infty} \frac{\partial \ln W(n_i)}{\partial n_i} \Delta n_i = 0 \tag{5.17}$$

We consider one term from the sum on the right-hand side. With Stirlings approximation $\ln x! \simeq x \ln x - x$ for large x we get

$$\begin{aligned} \ln W(n_i) &= (n_i + Z_i - 1)\ln(n_i + Z_i - 1) - (n_i + Z_i - 1) - \\ &\quad -n_i \ln n_i + n_i - (Z_i - 1)\ln(Z_i - 1) + (Z_i - 1) = \\ &= (n_i + Z_i - 1)\ln(n_i + Z_i - 1) - n_i \ln n_i - (Z_i - 1)\ln(Z_i - 1) \end{aligned} \tag{5.18}$$

For a nearby number $n_i + \Delta n_i$, we have

$$\begin{aligned} \Delta \ln W(n_i) &\equiv \ln W(n_i + \Delta n_i) - \ln W(n_i) \simeq \Delta n_i \frac{\partial \ln W(n_i)}{\partial n_i} \\ &= \Delta n_i \left[\ln(n_i + Z_i - 1) - \ln n_i\right] \end{aligned} \tag{5.19}$$

to *first order* in Δn_i.
For equilibrium, i.e. the most likely particle distribution, we now have to set:

$$\Delta \ln W = \sum_{i=1}^{\infty} \Delta n_i \left[\ln(n_i + Z_i - 1) - \ln n_i\right] = 0 \tag{5.20}$$

Since we consider a system in thermodynamic equilibrium, i.e. for which the number of particles $N = \sum n_i$ per unit volume and the energy $E = \sum n_i \epsilon_i$ per unit volume are constant, the variations in n_i must be such as to conserve N and E i.e.

$$\Delta N = \sum_{i=1}^{\infty} \Delta n_i = 0 \tag{5.21}$$

and

$$\Delta E = \sum_{i=1}^{\infty} \epsilon_i \Delta n_i = 0 \tag{5.22}$$

These restrictions imply that we also have:

$$\Delta \ln W - \alpha \Delta N - \beta \Delta E \;=\; \sum_{i=1}^{\infty} \Delta n_i \left[\ln(n_i + Z_i - 1) - \ln n_i - \alpha - \beta \epsilon_i\right] \;=\; 0 \tag{5.23}$$

The sum in (5.23) is zero for arbitrary variations Δn_i, provided that for each i

$$\ln(\overline{n}_i + Z_i - 1) - \ln \overline{n}_i - \alpha - \beta \epsilon_i = 0 \Rightarrow \frac{\overline{n}_i}{Z_i - 1} = \frac{1}{e^{\alpha + \beta \epsilon_i} - 1} \tag{5.24}$$

which is the Bose Einstein distribution.
Since $Z_i \gg 1$, $\overline{n}_i/(Z_i - 1)$ can be replaced by $\overline{n}_i/Z_i$, which represents the average occupation at energy level ϵ_i (occupation number). The actual values of α and β depend on the total number of particles and the total energy, and can be determined from these by substituting n_i in $N = \sum_{i=1}^{\infty} n_i$ and in $E = \sum_{i=1}^{\infty} n_i \epsilon_i$. For the Planck function, the number of *photons* need not be conserved (an atom can absorb a photon and jump from orbit 1 to orbit 3, and then emit 2 photons by returning via orbit 2)!. Thus, the *Lagrange condition* (5.21) does not apply to photons, and we obtain the Planck function for photons by dropping α in (5.24):

$$\frac{\overline{n}_i}{Z_i} = \overline{n}(\nu_k) = \frac{1}{e^{h\nu_k/kT} - 1} \tag{5.25}$$

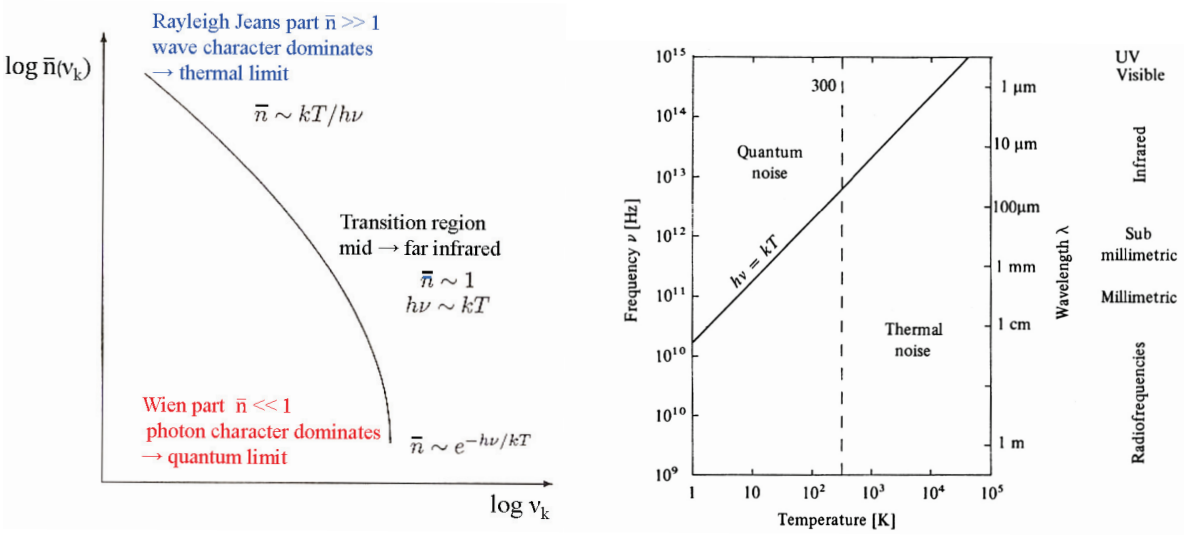


Figure 5.4: *Left: The mean occupation number $\bar{n}(\nu_k)$ ranging from $\ll 1$ to $\gg 1$ for a particular quantum level (photon state). The thermal limit (Rayleigh-Jeans part) and the quantum limit (Wien part) are both indicated. Right: Frequency as a function of temperature: division between thermal and quantum noise. Credit Lena et al. (1988)*

$\overline{n}(\nu_k)$ is the average occupation number of photons at energy $h\nu_k$, moreover we substituted $\beta = (kT)^{-1}$ that follows from thermodynamical considerations that we shall forego here. Note that photons do not collide directly with one another, but reach equilibrium only via interaction with atoms. If $h\nu_k \ll kT$, (5.25) reduces to $kT/h\nu_k \gg 1$, i.e. a large number of bosons that share the same quantum level (photon state). On the other hand if $h\nu_k \gg kT$, (5.25) becomes $e^{-h\nu_k/kT}$, i.e. the number of bosons that share the same quantum level is $\ll 1$. This situation is summarized in figure (5.4) that also links the various regimes of occupation numbers to the corresponding parts of the Planck distribution. For the *mean* energy (spectral power) equivalent of (5.25) we can write:

$$\overline{\epsilon}(\nu_k) \;=\; \frac{h\nu_k}{e^{h\nu_k/kT}-1} \tag{5.26}$$

This is the same expression that was used earlier in assigning the proper energy (spectral power) to a standing wave mode in (5.2).

Fluctuations around equilibrium

To investigate the fluctuations around equilibrium we shall look again at the number of ways in which $n_i + \Delta n_i$ photons can be distributed, as compared to that for n_i photons, where we now also include the second order term in the Taylor series development for $n_i + \Delta n_i$:

$$\ln W(n_i + \Delta n_i) = \ln W(n_i) + \Delta n_i \frac{\partial \ln W(n_i)}{\partial n_i} + \frac{\Delta n_i^2}{2}\frac{\partial^2 \ln W(n_i)}{\partial n_i^2} \tag{5.27}$$

In equilibrium the term in (5.27) containing the first derivative is by definition equal to zero. Consequently, we can rewrite 5.27 as:

$$W(n_i + \Delta n_i) = W(n_i)\mathrm{e}^{-\frac{W''(n_i)}{2}\Delta n_i^2} \qquad \text{where} \qquad W''(n_i) \equiv -\frac{\partial^2 \ln W(n_i)}{\partial n_i^2} \tag{5.28}$$

In other words, the probability of a deviation Δn_i drops exponentially with the square of Δn_i. The average value for Δn_i^2, the *variance* of the equilibrium distribution, is found by integrating over all values of Δn_i:

$$\overline{\Delta n_i^2} \;=\; \frac{\int_{-\infty}^{\infty} \Delta n_i^2 W(n_i) e^{-\frac{W''(n_i)}{2}\Delta n_i^2} d\Delta n_i}{\int_{-\infty}^{\infty} W(n_i) e^{-\frac{W''(n_i)}{2}\Delta n_i^2} d\Delta n_i} \;=\; \frac{1}{W''(n_i)} \tag{5.29}$$

Note that $W(n_i)$ and $W''(n_i)$ do not depend on Δn_i, so that they can be considered as constants in the integrations. *Note also* that the maximum negative deviation has $\Delta n_i = -n_i$, and the maximum positive deviation $\Delta n_i = N - n_i$; the integrals should therefore formally be evaluated between these values; however, for large Δn_i the integrand drops rapidly to zero, and so we can extend the integrals to the full range $-\infty$ to $+\infty$ without compromising the result.

Computing the second derivative of $\ln W(n_i)$ and changing sign we find for the variance of the equilibrium distribution of the photon gas:

$$\overline{\Delta n_i^2} = [W''(n_i)]^{-1} = \overline{n}_i \frac{(\overline{n}_i + Z_i - 1)}{Z_i - 1} = \overline{n}_i \left[1 + \frac{1}{e^{h\nu_k/kT} - 1}\right] = \overline{n_i}(1 + \overline{n}(\nu_k)) \quad (5.30)$$

The variance $\overline{\Delta n^2}(\nu_k)$ in the average occupation number follows from:

$$\overline{\Delta n^2}(\nu_k) \;=\; \frac{\overline{\Delta n_i^2}}{Z_i} \;=\; \overline{n}(\nu_k)(1 + \overline{n}(\nu_k)) \quad (5.31)$$

The volume density of photons in a blackbody Bose gas with frequencies between ν_k and $\nu_k + d\nu_k$ follows from $\bar{\phi}(\nu_k)d\nu_k = g(\nu_k)\bar{n}(\nu_k)d\nu_k$, in which $g(\nu_k)$ represents the volume density of quantum states per unit frequency at ν_k. The Bose-fluctuations $\overline{\Delta\phi^2}(\nu_k)$ in photon density per unit frequency can be written as (omitting the suffix k):

$$\overline{\Delta\phi^2}(\nu) = \left(\frac{8\pi}{c^3}\frac{\nu^2}{\exp\left(\frac{h\nu}{kT}\right) - 1}\right)\left(1 + \frac{1}{\exp(h\nu/kT) - 1}\right) \quad (5.32)$$

where we have used the expression for the equilibrium photon density $\bar{\phi}(\nu)$ of a black body photon gas.
If a detection element is placed within a black body radiation field inside a vacuum enclosure at temperature T, the incident photon flux is given by $\bar{n}(\nu) = \frac{1}{2}\frac{c}{4\pi}\bar{\phi}(\nu)A_e\Omega$. The factor $\frac{1}{2}$ refers to *one component* of polarization, A_e is the effective area of the detection element and Ω constitutes the solid angle subtended by the detector beam viewing the radiation field. If radiation illuminates an extended surface (A_e) with various directions of the wave vector, i.e. an omnidirectional black body radiation field, coherence theory states that spatial coherence is limited to $A_e\Omega \approx \lambda^2$, the so-called *extent of coherence*.[1] The *mean* photon flux $\bar{n}(\nu)$ (in photons $s^{-1}Hz^{-1}$) and its *variance* then becomes:

$$\bar{n}(\nu) \;=\; \frac{1}{\exp(h\nu/kT) - 1} \quad (5.33)$$

$$\overline{\Delta n^2}(\nu) \;=\; \bar{n}(\nu)\left(1 + \frac{1}{\exp(h\nu/kT) - 1}\right) = \bar{n}(\nu)\left[1 + \bar{n}(\nu)\right] \quad (5.34)$$

5.3.2 Thermal and quantum limits

The variance in the mean specific photon flux apparently contains two terms: one term proportional to $\bar{n}(\nu)$ and one term proportional to $\bar{n}^2(\nu)$.
For a photon energy $h\nu \ll kT$, we get $\bar{n}(\nu) = kT/h\nu$, but the noise power in this frequency regime is normally expressed in terms of the average radiation power $\bar{P}(\nu)$. We have then:

$$\bar{P}(\nu) = h\nu\bar{n}(\nu) = h\nu\frac{kT}{h\nu} = kT \qquad [\text{Watt Hz}^{-1}] \quad (5.35)$$

[1] This relation is the same as that governing the size $\theta = \lambda/D$ of a diffraction limited beam ($\Omega \approx \theta^2$) for an aperture with diameter D: $A_e \approx D^2$.

This is the expression for the *classical* thermal noise power per unit frequency bandwidth that we already encountered and derived in the previous section on thermal noise. Therefore, evidently, this limit of expression (5.33) is called the thermal limit and the noise power is independent of the frequency ν.
For the variance in $\bar{P}(\nu)$ we have:

$$\overline{\Delta P^2}(\nu) \;=\; \bar{P}^2(\nu), \quad \text{standard deviation} \quad \sqrt{\overline{\Delta P^2}(\nu)} \;=\; \bar{P}(\nu) \tag{5.36}$$

From this we can conclude that the fluctuations in the mean thermal noise power $\bar{P}(\nu)$ are of the same magnitude as the noise power itself. These fluctuations in the thermal limit signify that whenever wave packet interference becomes important, the interference will cause the fluctuations to become of the same magnitude as the signal. The low frequency fluctuations can be thought of as caused by the random phase differences and beats of the wavefield.
The transition between noise in the thermal limit into the quantum limit occurs at $h\nu \approx kT$. At room temperature, $T \approx 300$ K, this corresponds to a frequency $\nu \approx 6$ THz, or a wavelength $\lambda \approx 50$ μm. The relation $\nu = kT/h$ as a function of temperature T is displayed in figure (5.4). It is clear from this diagram that radio observations are always dominated by the wave character of the incoming beam and are therefore performed in the thermal limit. As a result, the treatment of noise in radio observations differs drastically from that of measurements at shorter wavelengths. Specifically at sub-millimeter and infrared wavelengths quantum limited observation is vigorously pursued but this remains still difficult.
In the extreme case that $h\nu \gg kT$, expression (5.33) reduces to $e^{-h\nu/kT} \ll 1$, hence:

$$\overline{\Delta n^2}(\nu) = \bar{n}(\nu) \tag{5.37}$$

This is the well-known expression for Poissonian noise in a sample containing $\bar{n}(\nu)$ photons. This condition implies the quantum limit of the fluctuations and it represents the minimum value of intrinsic noise present in any radiation field. In sensor physics and electronics this noise component is often referred to as *shot noise* and will be the subject of the next paragraph.

5.4 Shot noise

In contrast to thermal noise, shot noise involves an intrinsic fluctuation in electrical *current*, caused by an electronic component. Shot noise occurs when a discrete charge carrier (electrons and also holes in case of a semiconductor) crosses a potential barrier like an exit potential ('workfunction') in a photosensitive material, e.g. a photocathode, or by diffusion through a potential gradient in a semiconductor pn-junction. The underlying principal mechanism is that in the charge flow (= current) the passage of each individual charge carrier through a potential barrier occurs at random moments in time (t_i) that are completely independent of the other charge carriers in the flow. This causes irregularities in the carrier flow (i.e. carrier noise) that gives rise to a noisy electrical current $I(t) = I_0 + \Delta I(t)$, with I_0 the average value (= DC-component) and $\Delta I(t)$ the time varying noise component with average value zero.

5.4.1 The unfiltered Poisson process

The accumulating charge *flow* [number of charge impulses] represented by a time series of unfiltered random individual charge pulses can be described by a staircase function with discontinuities at time locations t_i, as shown in figure (5.5):

$$Z(t) = q\sum_i U(t-t_i),\ U(t) = \text{unit-step function},\ q = \text{elementary charge} \quad (5.38)$$

$$U(t) = \begin{cases} 1 & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

The charge *flow rate* [number of charge impulses per second] follows from time differ-

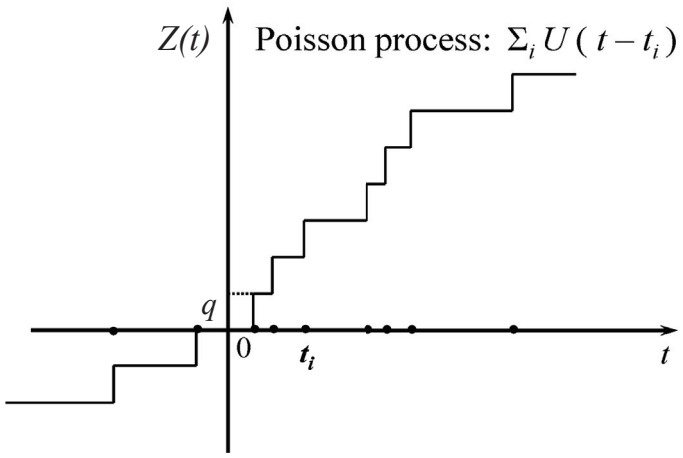


Figure 5.5: *Staircase function describing a Poisson process.*

entiation of the stochastic variable $Z(t)$:

$$X(t) = \frac{dZ(t)}{dt} = q\sum_i \delta(t-t_i) \quad (5.39)$$

and represents a train of Dirac impulses at random time locations t_i. At a constant average current I_0, $X(t)$ is a wide sense stationary (WSS) stochastic signal with a *time independent* average $\overline{X(t)} = q\lambda$ representing the amount of charge per unit time ($= I_0$), where λ is the *rate parameter* characteristic for the process under consideration.
We can now express the stochastic process $Z(t)$, displayed in figure (5.5), in the following way:

$$Z(t) = \int_0^t X(t')dt' = q\int_0^t \sum_i \delta(t'-t_i)dt' = qk(0,t) \quad (5.40)$$

in which $k(t_1,t_2)$ represents the *number* of charge impulses in a time period (t_1,t_2) of length $t = t_2 - t_1$. This number $k(t_1,t_2)$ is a Poisson distributed random variable (RV)

with parameter λt, i.e. $Z(t)$ expresses an unfiltered Poisson process:

$$\mathbf{p}\{k,\lambda t\} = \frac{(\lambda t)^k}{k!}e^{-\lambda t}, \quad \text{with } \lambda \text{ the rate parameter (see above)} \tag{5.41}$$

Note: For Poisson distributed RVs hold that if two time periods (t_1,t_2) and (t_3,t_4) are considered that are non-overlapping, then the RVs $k(t_1,t_2)$ and $k(t_3,t_4)$ are independent.
From expression (5.41) we can construct a Poissonian probability density function featuring a continuous random variable (κ):

$$\mathbf{p}(\kappa,\lambda t) = \sum_{k=0}^{\infty}\mathbf{p}(k,\lambda t)\delta(\kappa-k) \tag{5.42}$$

The average value of κ and of κ^2 for assessment of the fluctuation magnitude follow from:

$$\mathbf{E}\{\kappa\} = \int_{-\infty}^{+\infty}\kappa\,\mathbf{p}(\kappa,\lambda t)d\kappa = \lambda t \tag{5.43}$$

$$\mathbf{E}\{\kappa^2\} = \int_{-\infty}^{+\infty}\kappa^2\,\mathbf{p}(\kappa,\lambda t)d\kappa = (\lambda t)^2+\lambda t \tag{5.44}$$

The average value for $\kappa=\lambda t$ in equation (5.43) is of course as expected; the first term of equation (5.44) is the square of the average and its second term represents the variance. Since the variance of the fluctuations associated with the carrier noise equals λt, the standard deviation becomes $\sqrt{\lambda t}$, i.e. the 'strength' of the current noise. The relative fluctuation or signal to noise ratio (SNR) is then:

$$SNR = \frac{\lambda t}{\sqrt{\lambda t}} = \sqrt{\lambda t} \tag{5.45}$$

Consequently, the larger λt, the smaller the relative carrier noise. With very small λ we apparently need a long filter time to suppress the shot noise.
To determine the autocorrelation function $R_Z(t_1,t_2)$ of the Poisson process $Z(t)$ let us first consider $t_2 \geq t_1$. The variables $k(0,t_1)$ and $k(t_1,t_2)$, referring to adjacent but non-overlapping time periods, are then independent Poisson variables with parameters λt_1 and $\lambda(t_2-t_1)$ respectively. Thus we have:

$$\mathbf{E}\{k(0,t_1)k(t_1,t_2)\} = \mathbf{E}\{k(0,t_1)\}\mathbf{E}\{k(t_1,t_2)\} = \lambda^2 t_1(t_2-t_1), \text{ also } \Rightarrow \tag{5.46}$$

$$k(t_1,t_2) = k(0,t_2)-k(0,t_1) = \left[\frac{[Z(t_2)-Z(t_1)}{q}\right], \Rightarrow \text{ in (5.46)} \Rightarrow$$

$$\mathbf{E}\left\{\frac{Z(t_1)\,[Z(t_2)-Z(t_1)]}{q^2}\right\} = \left[\frac{R_Z(t_1,t_2)-\mathbf{E}\{Z^2(t_1)\}}{q^2}\right] \Rightarrow$$

$$R_Z(t_1,t_2) = q^2\left[\lambda^2 t_1(t_2-t_1)+\lambda^2 t_1^2+\lambda t_1\right] = q^2\left[\lambda^2 t_1 t_2+\lambda t_1\right] \tag{5.47}$$

$$\text{If } t_2<t_1 \Rightarrow R_Z(t_1,t_2) = q^2\left[\lambda^2 t_1 t_2+\lambda t_2\right] \tag{5.48}$$

Introducing the autocovariance $C_Z(t_1, t_2)$ of $Z(t)$ we can write:

$$R_Z(t_1, t_2) = q^2\lambda^2 t_1 t_2 + C_Z(t_1, t_2) = q^2\left\{\lambda^2 t_1 t_2 + \lambda t_1 U(t_2 - t_1) + \lambda t_2 U(t_1 - t_2)\right\} \quad (5.49)$$

Regarding the stochastic variable $X(t)$, the time derivative of $Z(t)$ and representing the train of Dirac impulses at random time locations, we have the time independent average value $\mathbf{E}\{X(t)\} = q\lambda$, with λ the rate parameter.
The autocorrelation function follows from successive partial differentiation of the autocorrelation of $Z(t)$ with respect to t_1 and t_2, thus:

$$R_X(t_1, t_2) = \frac{\partial^2 R_Z(t_1, t_2)}{\partial t_1 \partial t_2} = q^2\{\lambda + \delta(t_2 - t_1)\} \quad (5.50)$$

Designating the time difference $(t_2 - t_1) = \tau$ and dropping the normalization factor q^2 related to the *specific case* of charge impulses we arrive at the general expression for the autocorrelation of a train of unit-value Dirac impulses at random time positions (WSS ergodic signal):

$$R_X(\tau) = \lambda^2 + \lambda\delta(\tau) \quad (5.51)$$

The second term in equation (5.51) represents the covariance $C(\tau)$ of $X(t)$, which equals in this case the variance $C(0)$ since it is zero for every value of τ except for $\tau = 0$. This is of course evident, since the Dirac impulses are randomly distributed in time and are thus mutually completely uncorrelated.

5.4.2 Frequency limited shot noise

By applying the Wiener Khinchin theorem to $R_X(\tau)$ we can compute the power spectral density:

$$R_X(\tau) \Leftrightarrow S_{d_X}(\nu) = \int_{-\infty}^{+\infty} R_X(\tau) e^{-2\pi j\nu\tau} d\tau = \lambda^2\delta(\nu) + \lambda \quad (5.52)$$

which is inconsistent with physical reality since it implies an infinitely high power signal. In practice there is always a frequency cut-off at say ν_c, owing to some (high frequency) filtering process. We might perceive this by pinpointing that an RC network connected to a diode is triggered by each individual charge carrier that leaves the anode. The RC-network acts on each individual charge impuls q (δ-function) with a current response function $h(t)$. The resulting current in the anode circuit follows from a convolution of the Dirac δ-function train $X(t)$ with $h(t)$:

$$\frac{I(t)}{q} \;=\; X(t) \rightarrow h(t) \rightarrow Y(t) \quad (5.53)$$

with $h(t)$ the filter circuit *impulse* current response function (*Note*: $h(t) = 0$ for $t < 0$ and is a normalized function: $\int_0^\infty h(t)dt = 1$).

Hence we have:

$$Y(t) = h(t) * X(t) = \int_0^\infty \sum_k \delta(t' - t_k) h(t - t') dt' = \sum_k h(t - t_k) = \overline{Y} + \Delta Y(t) \quad (5.54)$$

Owing to the high carrier density in the charge flow, there will be a large degree of overlap between subsequent responses. This will result in a total current $I(t)$ that shows a Gaussian (normal) distribution around a mean value $\overline{I}$. For the expectation value of $Y(t)$ we thus find

$$\begin{aligned}\overline{Y} &= \mathbf{E}\{Y(t)\} = \mathbf{E}\{\int_0^\infty X(t-t')h(t')dt'\} \\ &= \int_0^\infty \mathbf{E}\{X(t-t')\}\, h(t')dt' = \lambda \int_0^\infty h(t')dt' = \lambda\, H(0) \end{aligned} \tag{5.55}$$

where for the last transition we have used:

$$\bar{H}(2\pi j\nu) \equiv \int_0^\infty h(t')e^{-2\pi j\nu t'}dt' \Rightarrow H(0) = \int_0^\infty h(t')dt' \tag{5.56}$$

In the Fourier domain we write for the *current* power spectral density:

$$\begin{aligned} S_{d_Y}(\nu) &= |\bar{H}(2\pi j\nu)|^2\, S_{d_X}(\nu) \\ &= \lambda^2 \left|\bar{H}(2\pi j\nu)\right|^2 \delta(\nu) + \lambda\, |\bar{H}(2\pi j\nu)|^2 = \lambda^2\, H^2(0) + \lambda\, |\bar{H}(2\pi j\nu)|^2 \end{aligned} \tag{5.57}$$

Evidently the power is now finite, as it should be. We obtain the autocorrelation by taking the Fourier transform of the *current* power spectral density $S_{d_Y}(\nu)$:

$$R_Y(\tau) = \lambda^2\, H^2(0) + \lambda\, [h(\tau) * h(\tau)] \tag{5.58}$$

where the first term on the right hand side gives the *mean* charge response of the linear dynamic system, and the second term represents the noise. Taking the autocovariance at $\tau = 0$ we obtain the variance of the noise signal:

$$C_Y(0) = \lambda \int_{-\infty}^{+\infty} h^2(t)dt = \lambda \int_{-\infty}^{+\infty} |\bar{H}(2\pi j\nu)|^2 d\nu = 2\lambda \int_0^{+\infty} |\bar{H}(2\pi j\nu)|^2 d\nu \tag{5.59}$$

in taking the last steps we have applied Parseval's theorem and changed from a double sided $S_{d_I}(-\infty < \nu < +\infty)$ to a one-sided S_I: twice the integral from $0 < \nu < \infty$ to accommodate physically real frequencies.
For the *mean square* output current $\overline{I^2}$ after filtering of the initial Poisson process, we can now write (multiplication by q^2):

$$\overline{I^2} = \overline{I}^2 + \overline{\Delta I^2} = [q\,\lambda\, H(0)]^2 + 2q^2\lambda \int_0^\infty |\bar{H}(2\pi j\nu)|^2 d\nu \quad [\text{Ampere}^2]$$

Substituting now the mean value of the initial, *unfiltered*, current source $I_0 = q\lambda \rightarrow$:

$$\overline{I^2} = [I_0\, H(0)]^2 + 2q\, I_0 \int_0^\infty |\bar{H}(2\pi j\nu)|^2 d\nu \quad [\text{Ampere}^2], \quad \text{with} \tag{5.60}$$

$$\overline{\Delta I^2} = 2q\, I_0 \int_0^\infty |\bar{H}(2\pi j\nu)|^2 d\nu = 2q\, I_0 \Delta\nu_c, \Delta\nu_c = \text{noise power bandwidth} \tag{5.61}$$

$$S_I(\nu) = 2q\, I_0\ [\text{Ampere}^2\,\text{Hz}^{-1}], \text{ the } \textit{current} \text{ spectral noise power } (\nu \ll \nu_c) \tag{5.62}$$

Clearly $S_I(\nu)$ is independent of frequency, i.e. the shot noise spectral noise power is 'white', its magnitude is contained by the *current* modulation transfer function $H(2\pi j\nu)$ of the associated filter circuitry. The above formula's are relevant for all shot noise processes in physics, apart from application in electronics they are also relevant for assessing clutter noise in radar images or for handling grain size noise in photographic plates.
As an example we derive here the variance of the *current* noise power by filtering with a simple RC-circuit. Each charge carrier fed into the RC filter will be immediately absorbed by the capacitor C and deposits an energy $q^2/2C$, the average power delivered equals $P = I_0 q/2C$. This power will eventually be dissipated in the resistor R, with a decay constant $\zeta = RC$. The normalized impulse response function $h(t)$ has an exponential decay with this time constant ζ, so we have:

$$h(t) = \frac{e^{-t/\zeta}}{\zeta} U(t) \Leftrightarrow \bar{H}(2\pi j\nu) = \frac{1}{\zeta}\int_0^\infty e^{-(\frac{1}{\zeta}+2\pi j\nu)t} dt = \frac{1}{1+2\pi j\nu\zeta} \Rightarrow$$

$$\overline{\Delta I^2} = 2q\, I_0 \int_0^\infty \frac{d\nu}{1+4\pi^2\nu^2\zeta^2} = \frac{qI_0}{\pi\zeta}\int_0^\infty \frac{dx}{1+x^2}\ (x = 2\pi\nu\zeta) = \frac{S_I(\nu)}{4RC} \tag{5.63}$$

The 'equivalent noise power bandwidth' imposed by the RC-filter equals $\Delta\nu = (4RC)^{-1} = (\pi/2)\nu_c$, which is slightly larger than the usual half power bandwidth $\nu_c = (2\pi RC)^{-1}$.
The power spectral density of shot noise $S_I(\nu) = 2qI_0$ can also be derived by applying a method similar to the one we used for the thermal noise power, i.e. by connecting a sharply tuned parallel LCR filter to the anode of a noisy diode. In saturation mode the diode can be regarded as a current source with a very high source impedance $R_s \rightarrow \infty$.The current of the saturated diode, with a *current* power spectral density $S_I(\nu)$, is fed into a parallel LCR network with a resonance frequency ω_0. This set up is shown in Figure (5.6).
The charge that is delivered into the capacitor generates an average power of $I_0\, q/2C$. This power is eventually dissipated in the resistor R for which we have a dissipative power $\overline{U^2}/R$, so we can write $\overline{U^2} = I_0\, qR/2C$. Next we can also express $\overline{U^2}$ in terms of the current spectral density $S_I(\nu)$ through:

$$\overline{U^2} = \int_0^\infty S_I(\nu)\, |\bar{Z}(j\omega)|^2\, d\nu = \int_0^\infty S_I(\nu) \left| \left\{ \frac{1}{R} + j(\omega C - \frac{1}{\omega L}) \right\}^{-1} \right|^2 d\nu \tag{5.64}$$

With ever increasing R the LCR circuit gets more sharply tuned at the resonance frequency $\omega_0 = 1/\sqrt{LC}$, until the only contribution to the integral is limited to a very small interval $\Delta\nu$ encompassing the resonance frequency ν_0. Introducing the same approximations as we used for the thermal noise case and implementing a change of variables (x), we can write:

$$\overline{U^2} = \frac{R\,S_I(\nu_0)}{4\pi C}\int_{-\infty}^{+\infty}\frac{dx}{1+x^2} \quad \text{with} \quad x = 2RC(\omega-\omega_0) \rightarrow = \frac{R\,S_I(\nu_0)}{4C} \tag{5.65}$$

(*Note*: For $\omega = 0 \rightarrow x = -2RC\omega_0$, with $R \rightarrow \infty,\ x \rightarrow -\infty$).
Equating the two results for $\overline{U^2}$ yields the wanted expression for the shot noise spectral density:

$$S_I(\nu) = 2qI_0 \quad [\text{Ampere}^2\text{Hz}^{-1}] \tag{5.66}$$

Maximum available noise spectral power, reduced shot noise

For the analysis of the maximum available shot noise power we refer here to the current source equivalent circuit displayed in Figure (2.1) and assume that all the complex impedances $\bar{Z}$ are resistive elements only, so we have a source resistor R_s in parallel with a load resistor R_L. For the power spectral density across the load resistor R_L in [Watt Hz^{-1}] we can write:

$$P(\nu) = \frac{S_V(\nu)}{R_L} = \frac{S_I(\nu)}{R_L}\left(\frac{R_sR_L}{R_s+R_L}\right)^2, \quad \text{maximize value} \Rightarrow \tag{5.67}$$

$$\frac{\partial P(\nu)}{\partial R_L} = 0 \Rightarrow 2qI_0R_s^2\frac{\partial}{\partial R_L}\left[\frac{R_L}{(R_s+R_L)^2}\right] = 0 \Rightarrow R_s = R_L \tag{5.68}$$

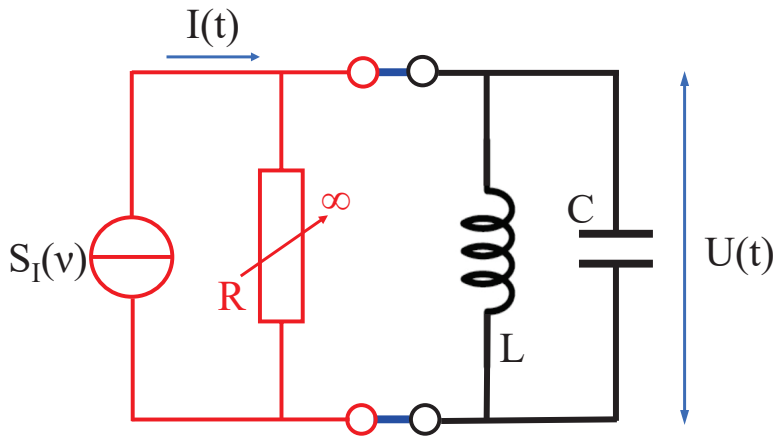


Figure 5.6: *A current noise source $I(t)$ feeding a parallel LCR circuit. The unknown current power spectral density distribution $S_I(\nu)$ can be assessed by selecting a large value for R (low circuit admittance) that gives rise to a sharp peak (small frequency interval) at the circuit's resonance frequency.*

(*Note*: this is the same result that we obtained earlier for the voltage equivalent circuit, for the case of complex impedances we get again $\bar{Z}_s^* = \bar{Z}_L$.)
Substitution in (5.67) yields the maximum spectral shot noise power that can be dissipated in the load R_L:

$$P_{shotnoise} = \frac{1}{2} q I_0 R_L \quad [\text{Watt Hz}^{-1}] \tag{5.69}$$

Obviously the load resistor will also produce thermal noise, so the maximum spectral noise power associated with e.g. a diode current load amounts to:

$$P_{max} = \left(\frac{1}{2} q I_0 R_L + kT \right) \quad [\text{Watt Hz}^{-1}] \tag{5.70}$$

Everything we have treated so far critically depends on the basic assumption that all the emission times t_k are strictly independent from each other. However in electronics mechanism may be at play that introduce interdependence in the charge carrier flow and, therefore, impair our basic assumption. This reduces the shot noise in the current $I(t)$ and is taken into account by adding a coefficient $r^2 < 1$ to the current power spectral density. i.e. $\overline{\Delta I^2} = 2r^2 q I_0$.
A classical example is the space charge effect in an unsaturated diode caused by to low a voltage on the anode to instantaneously absorb all the electrons in the current flow. If a number of electrons happen to be emitted shortly after each other than, momentarily, the space charge is slightly increased and electrons that are emitted somewhat later in the sequence will be pushed back to the cathode quickly compensating for the surplus of electrons that were emitted earlier. As a consequence the electron flow to the anode becomes slightly 'smoother' (less irregular) than in a pure Poisson process.
An extreme example of such a situation is the current flow through a conductor, for example a resistor or a piece of copper wire. This current does not exhibit any shot noise despite the fact that it is entirely composed of discrete charges. The explanation is found in the fact that a conductor cannot contain any stored charge (we neglect the parasitic capacitance) and, consequently, injection of an electron from the negative battery pole will be instantaneously compensated by the absorption of an electron by the positive battery pole. This mechanism forces the charge carrier flow to a very regular pattern in which shot noise is practically absent. The fact that a diode does exhibit shot noise is caused by the thermal emission by the cathode or by the diffusion of charge carriers through a pn-junction that involves crossing of a *potential barrier*, these charge carrier crossings are then mutually completely independent.

5.5 Generation-Recombination noise in semiconductors

Generation-Recombination noise (GR noise) originates in thermally or optically-stimulated electronic transitions between valance and conduction band or transitions between impurity levels, traps, or recombination centers and one of these bands. Associated with

these transitions are fluctuations in the numbers of free carriers and in their lifetimes, thus giving rise to the GR noise. The detailed mathematical treatment of GR noise depends on many specific parameters like the number of energy levels, the energies corresponding to these levels, the electron population, and the occupancy of states.
As a simple example, the common case of an extrinsic semiconductor such as Germanium or Silicon containing both donors and acceptors, one being predominant and exceeding in number the number of free carriers, will be treated here.
For a simple GR two-level system we assume a generation rate (number per unit time) $g(N)$ and a recombination rate $r(N)$ which describe the transition from the impurity level to the conduction band and the reverse (recombination) process. N is a random variable that represents the number of free carriers (predominantly electrons) in the conduction band at time t. We further assume that both rates g and r depend explicitly *only* on the momentaneous number of free carriers in the conduction band, $N(t)$. In general g is a decreasing function of N (i.e. negative slope), whereas r is an increasing function of N (i.e. positive slope). In the equilibrium (steady state) situation we have balance between the generation and recombination rates, say at a free carrier average number value $\overline{N} = N_e$ at time t. Hence, for $g_e = g(N_e)$ and $r_e = r(N_e)$ we have

$$g_e = r_e \quad \text{and } N \text{ normally distributed around } N_e \quad \Rightarrow \tag{5.71}$$

$$\mathbf{p}(N) = p(N_e)\exp -\frac{1}{2}\left[\frac{(N-N_e)^2}{\overline{\Delta N^2}}\right] \text{ with variance } \overline{\Delta N^2} = N_e \tag{5.72}$$

Taking the derivatives $g'_e = (dg/dN)_{N=N_e}$ and $r'_e = (dr/dN)_{N=N_e}$ as the generation rate and the recombination rate at the equilibrium number value N_e respectively (dimension $[\sec^{-1}]$), we can assign specific time scales to the GR process by defining $1/\tau_g = -g'_e$ and $1/\tau_r = r'_e$ leading to a free carrier life time:

$$\frac{1}{\tau_\ell} = -\frac{d}{dN}(g-r)_{N=N_e} \quad \Rightarrow \quad \frac{1}{\tau_\ell} = -(g'_e - r'_e) = \frac{1}{\tau_g} + \frac{1}{\tau_r} \text{ and:} \tag{5.73}$$

$$\overline{\Delta N^2} = N_e = g_e \cdot \tau_\ell \tag{5.74}$$

The carrier life time τ_ℓ dictates the dynamical response of the semiconductor on changes in free carrier generation, this response can be quantified by solving the time dependent continuity equation for change $dN(t)/dt$:

$$\frac{dN(t)}{dt} = g - \frac{N(t)}{\tau_\ell} \quad \Rightarrow \quad N(t) = g\tau_\ell\left(1 - e^{\frac{-t}{\tau_\ell}}\right) \tag{5.75}$$

The dynamical behavior expressed in equation (5.75) is characterized by a first order system with transfer function $H(2\pi j\nu) = 1/(1 + 2\pi j\nu\tau_\ell)$, its frequency response $|H(2\pi j\nu)|$ trails off at high frequencies with a tipping point at $\nu_\ell = 1/(2\pi\tau_\ell)$. This is shown, one-sided, in the left panel of figure (5.7). The associated autocovariance function is shown in the right panel of figure (5.7). It constitutes a double-sided exponential function centered on $\tau = 0$ with a decay constant τ_ℓ that follows from the

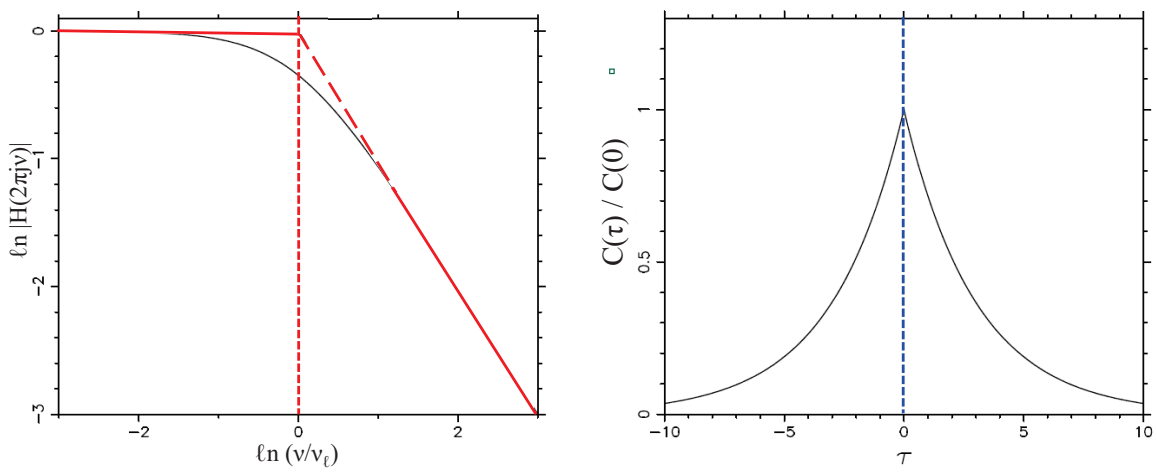


Figure 5.7: *For a first order transfer function with tipping point $\nu = \nu_\ell$ (left), the autocovariance drops exponentially with $|\tau|$ (right).*

Fourier transform of the *double sided* transfer function:

$$\begin{aligned} H(2\pi j\nu) + H(-2\pi j\nu) &= \left(\frac{1}{(1+2\pi j\nu\tau_\ell)} + \frac{1}{(1-2\pi j\nu\tau_\ell)}\right) \Rightarrow \text{FT} \Rightarrow \\ &\Leftrightarrow \frac{e^{-\frac{\tau}{\tau_\ell}}}{2\tau_\ell}U(\tau) + \frac{e^{\frac{\tau}{\tau_\ell}}}{2\tau_\ell}U(-\tau) = \\ &= \frac{e^{-\frac{|\tau|}{\tau_\ell}}}{2\tau_\ell} = C(\tau) \end{aligned} \tag{5.76}$$

where $U(\tau)$ is the Heaviside step function: $U(\tau) = 0$, $\tau < 0$; $U(\tau) = 1$, $\tau \geq 0$. Hence we have:

$$C(\tau) = C(0)\, e^{-\frac{|\tau|}{\tau_\ell}} \quad \text{with} \quad \tau_\ell = \frac{1}{2\pi\nu_\ell} \tag{5.77}$$

For the GR-process, featuring the random count variable N, we can thus express the variance $\overline{\Delta N^2}$ for some time delay τ following excitation *or* decay as an exponential autocovariance according to:

$$C_N(\tau) = \overline{\Delta N^2}\, e^{-\frac{|\tau|}{\tau_\ell}} \tag{5.78}$$

The associated power spectral density $S_N(\nu)$ can now be obtained by applying the Wiener-Khinchin theorem. Since we are dealing here with two *independent* random processes, i.e. an excitation process followed by a decay process, in computing the spectral noise power we incorporate a factor 2 in taking the integral of the autocovariance over all physical delays τ. Subsequently we need to convert this *one-sided* spectral density to a *double-sided* spectral density (S_{d_N}) to accommodate the *negative*

frequencies and time delays used in Wiener-Khinchin theorem. Thus we have:

$$S_N(\nu) = 2\int_0^{\infty} \overline{\Delta N^2}\, e^{-\frac{|\tau|}{\tau_\ell}}\, e^{-2\pi j\nu\tau} d\tau \quad \Rightarrow \quad S_{d_N}(\nu) = \int_{-\infty}^{\infty} \overline{\Delta N^2}\, e^{-\frac{|\tau|}{\tau_\ell}}\, e^{-2\pi j\nu\tau} d\tau \tag{5.79}$$

Performing the Fourier transform in equation (5.79) results in:

$$S_{d_N}(\nu) = \frac{2\tau_\ell \overline{\Delta N^2}}{1+(2\pi\nu\tau_\ell)^2} = \frac{2g_e\tau_\ell^2}{1+(2\pi\nu\tau_\ell)^2} \tag{5.80}$$

As shown before, the value of the average current density $|\vec{j}|$ in the semiconductor equals $nq|\vec{v}_d|$ with $n = N/V$ the charge carrier volume density, q the elementary charge and $\vec{v}_d$ the drift velocity in the applied electric field. With a cross sectional area A we have a total average current $I_e = A \cdot |\vec{j}| = A \cdot (N_e/V)q(d/\tau_{tr}) = qN_e/\tau_{tr} = qg_e(\tau_\ell/\tau_{tr})$ in which d represents the distance between the electrodes of the semiconductor and $\tau_{tr} = d^2/(\mu V)$ the charge carrier transit time between the electrodes (with μ the carrier mobility and V the bias voltage). Substituting in (5.80) and multiplying $S_{d_N}(\nu)$ by $(q/\tau_{tr})^2$ yields an expression for the current spectral density of the GR noise:

$$S_{d_I} = \left(\frac{q}{\tau_{tr}}\right)^2 S_{d_N}(\nu) = 2qI_e\left(\frac{\tau_\ell}{\tau_{tr}}\right)\left(\frac{1}{1+(2\pi\nu\tau_\ell)^2}\right) \tag{5.81}$$

The mean square GR current noise $\overline{\Delta I^2}$ follows from integration of S_{d_I} over all frequencies:

$$\overline{\Delta I^2} = 2qI_e\left(\frac{\tau_\ell}{\tau_{tr}}\right)\int_{-\infty}^{+\infty} \frac{d\nu}{1+(2\pi\nu\tau_\ell)^2} \Rightarrow \quad \overline{\Delta I^2} = 4qI_e\left(\frac{\tau_\ell}{\tau_{tr}}\right)\Delta\nu_c \text{ with} \tag{5.82}$$

$$\Delta\nu_c = \int_0^{+\infty} \frac{d\nu}{1+(2\pi\nu\tau_\ell)^2} \text{ the noise equivalent bandwidth within } 0 < \nu < \infty$$

For low frequencies the (*one-sided*) current spectral power can be expressed as:

$$S_I(0) = \frac{\overline{\Delta I^2}}{\Delta\nu_c} = 4qG_nI_e \quad [\text{Ampere}^{\,2}\ \text{Hz}^{\,-1}] \quad \text{with} \quad G_n = \left(\frac{\tau_\ell}{\tau_{tr}}\right) = \frac{\tau_\ell\mu E}{d} \tag{5.83}$$

G_n is the so-called *noise gain*. In case the semiconductor has uniform resistance and an uniform electric field, the noise gain is proportional to this applied electric field. The GR-current noise can then be expressed as $\left(\sqrt{\overline{\Delta I^2}}\right)_{GRn} = \sqrt{4qG_nI_e\Delta\nu_c}$, with $\left(\sqrt{\overline{\Delta I^2}}\right)_{GRn}$ the rms-noise current, I_e the average total current and $\Delta\nu_c$ the noise equivalent bandwidth. In case the extrinsic bulk semiconductor is employed as photoconductor, the photo-current equals $I_p = q\eta\Phi_{ph}G_p$, with Φ_{ph} the radiant signal photon flux, η the quantum efficiency of photoabsorption and G_p the *photoconductive gain*. Since in many photo conductive detectors GR-noise is the dominant noise source in

the frequency range of interest, the experimental method usually employed to obtain the value of the photoconductive gain relies on current noise measurements. It is then obviously assumed that both gains are equal $G_p = G_n$. It should however be born in mind that this *principally assumes a uniform generation of excess charge carriers* during exposure by a radiation source.
Expression (5.83) for the GR power spectral density resembles that for the classical shot noise spectral power derived earlier, see equation (5.62) for e.g. a thermionic diode, albeit with a different numerical factor (4 instead of 2) and an additional gain factor G. Although the origin of GR-noise is the fluctuations in the rates of generation and recombination of charge carriers and the origin of true shot noise is the corpuscular nature of these carriers, associated with emission or conduction, GR noise can become shot noise in the limit that the transit time t_{tr} becomes the shortest time scale as compared to the life times in the GR process. This condition actually implies that the total number of carriers inside the statistical sample fluctuates in time solely because of the random entering and exiting process of the carriers from the the electrical contacts. The low-frequency spectral density of the current fluctuations, covering continuously the transition from GR (i.e. $\tau_{tr} \gg \tau_\ell$) to shot noise (i.e. $\tau_{tr} \ll \tau_\ell$), can be described with:

$$S_I(0) = 4qI_e \left(\frac{\tau_\ell}{\tau_{tr}}\right) \left\{1 + \frac{\tau_\ell}{\tau_{tr}} \left[e^{-(\tau_{tr}/\tau_\ell)} - 1\right]\right\} \quad \text{yielding} \tag{5.84}$$

$$S_I(0) = 4qGI_e \quad \text{for} \quad \tau_{tr} \gg \tau_\ell \quad \text{and} \quad S_I(0) = 2qI_e \quad \text{for} \quad \tau_{tr} \ll \tau_\ell \tag{5.85}$$

Equation (5.83) relates to extrinsic GR noise in a biased, bulk semiconductor.
Let us also briefly consider a zero-biased pn junction. In the bulk regions, which are the regions adjacent to the depletion region, the electric field is of negligible strength. Hence, drift is inconsequential and diffusion is the relevant transport mechanism. Of the free carriers in the bulk regions, only minority carriers within a diffusion length L of the junction region are able to contribute to the junction noise and, as a consequence, the appropriate time of transport τ_{tr} is the lifetime τ_ℓ. Those carriers reaching the depletion region boundary are swept across the junction region (by the built-in field), and "instantaneously" collected with virtually no recombination having occurred. For the GR-noise of a zero biased pn-junction we thus have, since $\tau_{tr} = \tau_\ell$, the following expression for the low frequency spectral noise power:

$$S_I(0) = 4qI_e \quad [\text{Ampere}^{\,2}\ \text{Hz}^{\,-1}] \tag{5.86}$$

This expression for GR-noise in a pn-junction is also quite similar to that for the shot noise spectral power apart from the numerical constant, however it lacks the noise gain factor G_n. The presence of this gain factor is a distinct difference between photoconducting devices and photodiodes regarding their ultimate limiting sensitivity for light detection.

5.6 Phonon (temperature) noise

Consider a thermal sensor that can be regarded as a thermodynamically closed system of fixed volume, in thermal equilibrium with a heat bath. Its thermal energy states

can then be described with the aid of the canonical ensemble that gives the probabilities of the various possible energy states of such a closed system: it has a probability distribution with the Boltzmann form. The Boltzmann distribution is a probability distribution that gives the probability of a certain state as a function of that state's energy and temperature of the system to which the distribution is applied. The probability $\mathbf{p}(\epsilon_i) = \mathbf{p}_i$ that the system has a defined quantum state with energy ϵ_i is given by

$$\mathbf{p}_i = \frac{e^{-\epsilon_i/(kT)}}{J} \quad \text{with} \quad J = \sum_i e^{-\epsilon_i/(kT)} \tag{5.87}$$

The value of the coeffient J follows from the normalization requirement that the integral probability taken over all energy levels ϵ_i should be equal to unity. This expression implicitly assumes that the number of particles N in the canonical ensemble remains constant.
The fluctuation in the energy state ϵ can now be obtained by computing the variance $\overline{(\Delta\epsilon)^2} = \overline{\epsilon^2} - \bar{\epsilon}^2$ by applying the probability distribution function (5.87):

$$\bar{\epsilon} = \frac{\sum_i \epsilon_i e^{-\epsilon_i/(kT)}}{J} \quad \text{and} \quad \overline{\epsilon^2} = \frac{\sum_i \epsilon_i^2 e^{-\epsilon_i/(kT)}}{J} \tag{5.88}$$

Substituting for the moment $\beta = 1/(kT)$ we have:

$$\begin{aligned} \sum_i \epsilon_i e^{-\epsilon_i/(kT)} &= -\frac{dJ}{d\beta} \\ \sum_i \epsilon_i^2 e^{-\epsilon_i/(kT)} &= \frac{d^2J}{d\beta^2} \end{aligned}$$

This yields for the variance of the energy ϵ the following expression:

$$\begin{aligned} \overline{(\Delta\epsilon)^2} &= \overline{\epsilon^2} - \bar{\epsilon}^2 = \frac{1}{J}\frac{d^2J}{d\beta^2} - \frac{1}{J^2}\left(\frac{dJ}{d\beta}\right)^2 = \frac{d}{d\beta}\left(\frac{1}{J}\frac{dJ}{d\beta}\right) = \\ &= -\frac{d\bar{\epsilon}}{d\beta} = kT^2\frac{d\bar{\epsilon}}{dT} = kT^2C, \; C = \text{heat capacity of thermal sensor} \end{aligned} \tag{5.89}$$

The rms-energy fluctuation of the thermal phonon noise $\sqrt{\overline{(\Delta\epsilon)^2}} = \sqrt{kT^2C}$ determines the ultimate energy resolution that can be achieved with any thermal detector.
From the magnitude of the fluctuation in thermal energy we can derive the variance of the temperature fluctuations $\overline{\Delta T^2}$:

$$\overline{(\Delta\epsilon)^2} = C^2\overline{\Delta T^2} \quad \Rightarrow \quad \overline{\Delta T^2} = \frac{kT^2}{C} \tag{5.90}$$

Next, we can assess the frequency characteristics of the phonon noise by deriving an expression for its power spectral density. The frequency response of a thermal sensor is determined by the heat capacity C [Joule/K] of the sensing material and the thermal conductance G [Watt/K] from the sensor to the ambient temperature reference normally

embodied by a heat bath. We can write de following heat balance equation (first order system) for the temperature rise ΔT of the sensor in response to a heat impulse [Joule] $\int_{-\infty}^{+\infty} \Phi_{th}(t)\delta(t)dt$ with $\Phi(t)$ in [Watt]:

$$\begin{aligned} C\frac{d\Delta T}{dt} + G\Delta T &= \int_{-\infty}^{+\infty} \Phi_{th}(t)\delta(t)dt \quad \text{which can be written as:} \\ C\frac{d\Delta T}{dt} + G\Delta T &= 0 \text{ for } t > 0 \text{ and } C\frac{d\Delta T}{dt} + G\Delta T = \Phi_{th}(0) \text{ for } t = 0, \text{ with solution:} \\ \Delta T &= \frac{\Phi_{th}(0)}{G} e^{-t/\xi} \text{ [K]} \quad \text{with } \xi = \frac{C}{G} \Rightarrow \text{ intrinsic frequency cut-off} \end{aligned} \tag{5.91}$$

Making the *a priori* assumption that the phonon noise exhibits, in good approximation, a white spectrum until the intrinsic frequency limitation sets in, we can write the (two sided) power spectral density as $\eta|H(2\pi j\nu)|^2$ [K^{2}Hz^{-1}]. The value of η can now be obtained from the temperature variance derived above:

$$\overline{\Delta T^2} = \frac{kT^2}{C} = \int_{-\infty}^{+\infty} S_{d_{\Delta T}}(\nu)d\nu = \eta \int_{-\infty}^{+\infty} |\bar{H}(2\pi j\nu)|^2 d\nu = 2\eta \int_{0}^{+\infty} |\bar{H}(2\pi j\nu)|^2 d\nu \tag{5.92}$$

in taking the last step we have changed from a double sided ($S_{d_{\Delta T}}(\nu) \Rightarrow -\infty < \nu < +\infty$) to a one-sided spectrum $S_{\Delta T}(\nu)$: twice the integral from $0 < \nu < +\infty$.
Substituting $|\bar{H}(2\pi j\nu)|^2$ we get:

$$\begin{aligned} \frac{kT^2}{C} &= \frac{\eta}{\pi\xi} \int_0^{\infty} \frac{dx}{1+x^2} \; (x = 2\pi\xi\nu) \Rightarrow \eta = \frac{2kT^2}{G} \Rightarrow \overline{\Delta T^2} = \frac{4kT^2}{G}\Delta\nu_c, \; \Delta\nu_c = \frac{1}{4\xi} \Rightarrow \\ S_{\Delta T}(\nu) &= \frac{4kT^2}{G} \text{ [K}^2\text{Hz}^{-1}\text{] and } S_{\Delta P_{th}}(\nu) = G^2 S_{\Delta T}(\nu) = 4kT^2 G \text{ [Watt}^2\text{Hz}^{-1}\text{]} \end{aligned} \tag{5.93}$$

with $S_{\Delta T}(\nu)$ and $S_{\Delta P_{th}}(\nu)$ the one-sided power spectral densities of the temperature T and the thermal power P_{th} fluctuations for $\nu \ll 1/(2\pi\xi)$ and $\Delta\nu_c$ the equivalent noise power bandwidth.

5.7 Other noise sources

Very briefly we mention here three more sources of noise in electronics.

5.7.1 Partition noise

In a transistor, the emitter current is divided over the base and the collector, in a pentode vacuum tube the cathode current subdivides itself over the anode and the shielding grid and so on. This implies that the original charge carrier number density λ, see earlier, will be lowered to λ' in electrode #1 and to $\lambda - \lambda'$ in electrode #2. This

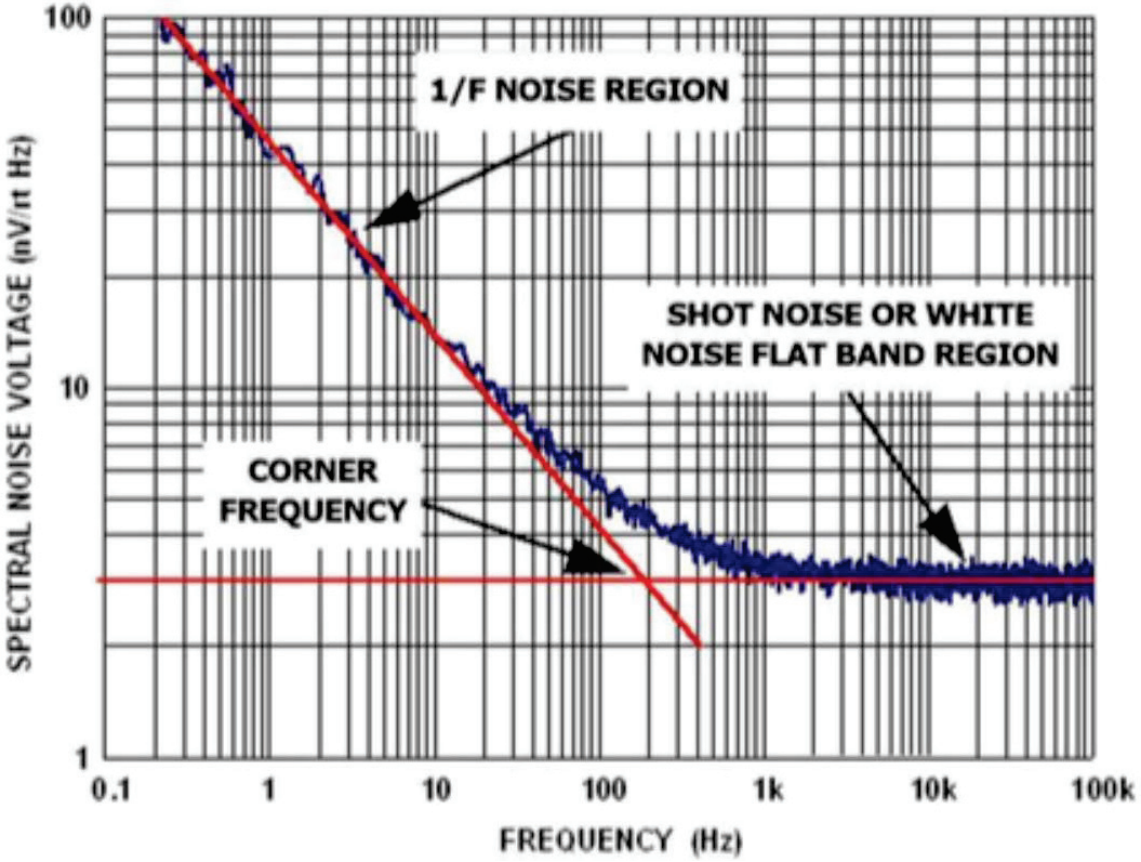


Figure 5.8: *Example of spectral noise voltages [Volt $Hz^{-1/2}$] as a function of frequency for 1/f−noise and white noise source components (thermal and shot noise). Below 1 kHz 1/f−noise clearly dominates, above several kHz the 1/f component becomes negligible and the white noise sources rule the noise factor.*

means a charge carrier loss, i.e. $\lambda' < \lambda$, for electrode #1 causing an increase of the relative carrier noise, see equation (5.45). The same happens when we consider the case for electrode #2. *Note:* Beware that in this partition case the fluctuations in each individual electrode are *correlated*. So theoretically one should be able to compensate the fluctuations in electrode #1 by those occurring in electrode #2.

5.7.2 1/f−noise

In diodes, transistors, charge emitting cathodes and also carbon-based resistors, the charge carrier transport can be concentrated in small regions of the conducting material that may fluctuate randomly in location and size. These fluctuations in the conductivity process introduce a current noise with a power spectral density:

$$S_I(\nu) \;=\; a \cdot \frac{I_0^n}{\nu} \qquad \text{with n} \leq 2, \tag{5.94}$$

hence inversely proportional to the frequency and proportional to the square of the average current I_0. This so-called 1/f−noise may dominate at low frequencies over the

thermal and shot noise components see figure(5.8), but tends to become negligible at higher frequencies, i.e. $\geq$ 10 kHz $\rightarrow$ 1 MHz, depending on the type of semiconductor utilized in the circuitry.

5.7.3 Microphonic noise

Microphonic noise is caused by the mechanical displacement of wiring and components when the sensor system is subjected to vibration and/or shock. Changes in capacitance following a change in wire spacing relative to the grounded reference planes (e.g. metal casing) is the major source responsible for this noise component. This is obviously very dependent on the sensor configuration details and cannot be modeled with a general expression. Operationally a sensor system should be constructed in a mechanically stable way, if a modulation of the sensor output does occur during a mechanical shock, a redesign of the mechanical arrangement of the sensor components is mandatory.

Chapter 6

Radiation sensors: generic aspects

6.1 Radiometry: the main radiometric definitions

In characterizing a radiation beam, the following main radiometric definitions can be distinguished:

- Radiant energy: Q, dimension [Joule].
  Note: energy of electromagnetic radiation.
- Radiant flux (Radiant power): $\Phi = dQ/dt$, dimension [Joule $\sec^{-1}$ = Watt].
  Note: radiant energy emitted, reflected, transmitted or received per unit time.
- **Irradiance (Radiant flux density)**: $I = d\Phi/dA$, dimension [Watt m^{-2}].
  Note: radiant flux received by a *surface* per unit area (sometimes referred to as 'intensity').
- **Radiance**:

$$L = \frac{dI}{\cos\theta\ d\Omega} = \frac{dQ}{\vec{r}\cdot\vec{n}\ dA\ d\Omega\ dt} \quad \text{dimension [Watt m}^{-2}\text{ sr}^{-1}\text{]} \quad (6.1)$$

$$\text{or} \quad dQ = L\ \vec{r}\cdot\vec{n}\ dA\ d\Omega\ dt,$$

  with L being the basic quantity describing a diffuse radiation field, $\vec{n}$ the unit vector normal to the area dA and $\vec{r}$ the unit direction vector pointing in the direction of the incident radiation.
  Note: radiant flux emitted, reflected. transmitted or received by a *surface* per unit solid angle per unit *projected* area. This is a *directional* quantity (also sometimes referred to as 'intensity').

- Spectral Radiance:

$$L_\lambda = \frac{dQ}{\vec{r}\cdot\vec{n}\ dA\ d\Omega\ d\lambda\ dt} \quad \text{dimension [Watt m}^{-2}\text{ sr}^{-1}\text{nm}^{-1}\text{]} \quad (6.2)$$

$$L_\nu = \frac{dQ}{\vec{r}\cdot\vec{n}\ dA\ d\Omega\ d\nu\ dt} \quad \text{dimension [Watt m}^{-2}\text{ sr}^{-1}\text{Hz}^{-1}\text{]} \quad (6.3)$$

$$L_\epsilon = \frac{dQ}{\vec{r}\cdot\vec{n}\ dA\ d\Omega\ d\epsilon\ dt} \quad \text{dimension [Watt m}^{-2}\text{ sr}^{-1}\text{eV}^{-1}\text{]} \quad (6.4)$$

Note: the spectral radiance incorporates the normalization of the radiance per unit wavelength or frequency/energy bandwidth. The monochromatic radiances are labeled as L_λ, L_ν or L_ϵ respectively, normalization is usually done per one-nanometer wavelength interval, one-Hz frequency interval or one-electron-Volt ($= 1.6 \cdot 10^{-19}$ Joule) energy interval.
In some disciplines (e.g. astronomy & astrophysics) the spectral radiance is (somewhat confusingly) referred to as *spectral* (or *specific*) 'intensity'.

6.2 Reflectance and transmittance at the optical interface with a sensing element

6.2.1 Snell's and Fresnel's laws

The interplay of a radiation beam with the entrance surface (interface) of a sensing element can be subdivided in a number of physical processes:

- **Reflection**
- **Refraction**
- **Transmission** and **Absorption**
- **Diffraction** and **Interference**
- **Polarization**

Elementary optical theory shows that the incident, reflected and transmitted rays at the interface between to media with refractive indices n_i and n_t are geometrically related in the following way:

- all rays are located in a single flat plane perpendicular to the sensor surface
- given an angle of incidence θ_i and an angle of reflection θ_r, both relative to the local vertical, $\theta_i = \theta_r$
- given an angle θ_t with the local vertical for a transmitted ray, θ_t and θ_i are related through Snell's law $n_i \sin\theta_i = n_t \sin\theta_t$

The *amplitude* and the *irradiance* of the reflected and transmitted beam at the interface are governed by Fresnel's laws, consequently these laws play an important role by the selection of the appropriate boundary layer between the sensor and the surrounding medium (e.g. the attainable detection efficiency). Hence, a short summary of the relevant equations is given in the next paragraph.
Consider a plane monochromatic wave with an electric vector $\vec{E}_i$ (amplitude $\vec{E}_{0i}$) incident on a planar surface separating two homogeneous and isotropic media:

$$\vec{E}_i = \vec{E}_{0i}\, e^{i(\vec{k}_i \cdot \vec{r} - \omega_i t)} \qquad \text{or} \qquad \vec{E}_i = \vec{E}_{0i}\, \cos(\vec{k}_i \cdot \vec{r} - \omega_i t) \tag{6.5}$$

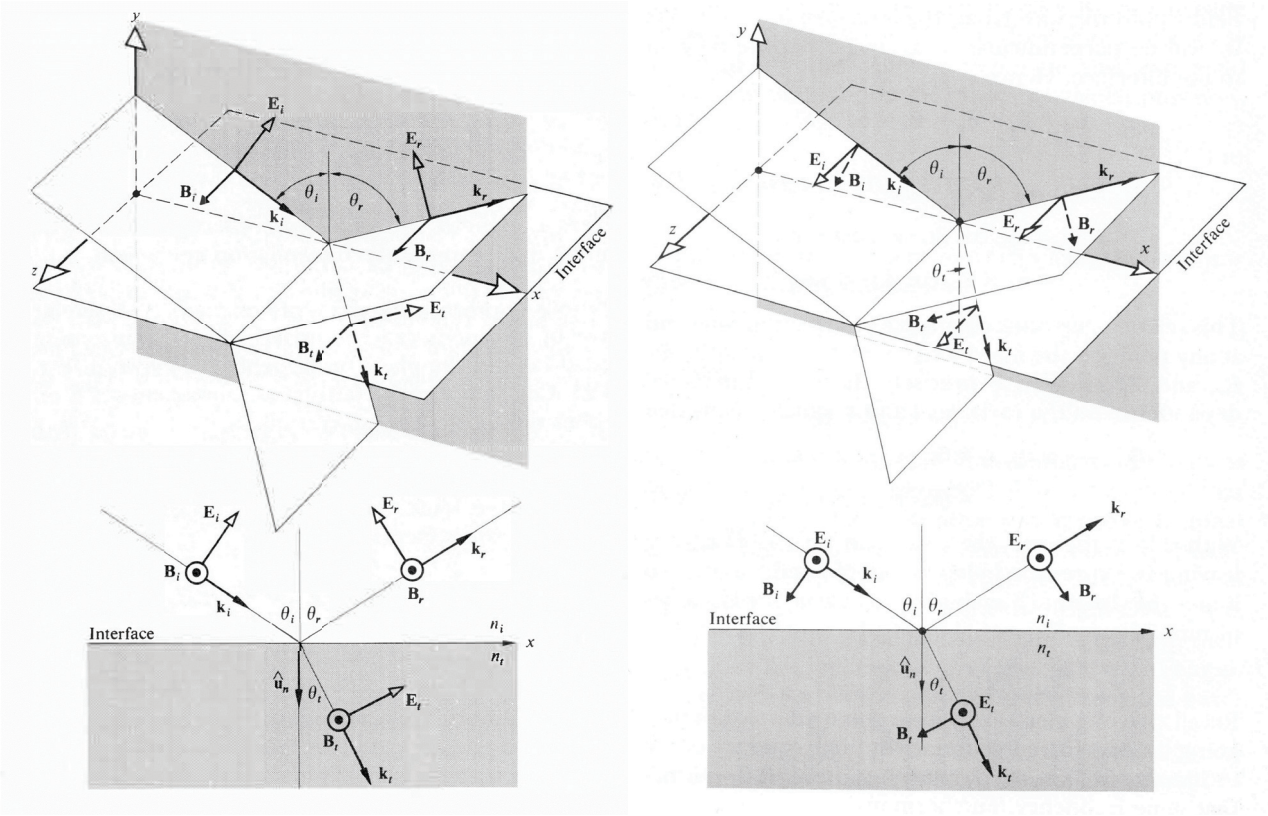


Figure 6.1: *Incoming EM-waves whose $\vec{E}$-fields are in the plane of incidence(left) and normal to the plane of incidence(right) respectively, credit Hecht(1987).*

with $\vec{k}_i$ the wave vector of the incident wave, $\vec{r}$ the wave position vector and ω_i the wave radial frequency.
Assume that $\vec{E}_{0i}$ is constant in time, that is the wave is linearly or plane polarized. Any form of EM-radiation can be represented by two orthogonal linearly polarized waves, so this does not constitute a restriction. The reflected and transmitted waves can be represented by $\vec{E}_r$ (amplitude $\vec{E}_{0r}$) and $\vec{E}_t$ (amplitude $\vec{E}_{0t}$) respectively. All the amplitude vectors can be split into two vector components: one perpendicular to the plane of incidence, $\left(\vec{E}_{0i,r,t}\right)_\perp$ and the other component $\left(\vec{E}_{0i,r,t}\right)_\parallel$ parallel to the plane of incidence, see Figure 6.1.
Fresnel's equations state that the amplitude *absolute values* of the reflected components $|\vec{E}_{0r}|_\perp$ and $|\vec{E}_{0r}|_\parallel$, normalized to the incident components can be expressed as:

$$\frac{|\vec{E}_{0r}|_\perp}{|\vec{E}_{0i}|_\perp} = r_\perp = -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)} \quad \text{and} \quad \frac{|\vec{E}_{0r}|_\parallel}{|\vec{E}_{0i}|_\parallel} = r_\parallel = +\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)} \tag{6.6}$$

In a similar way the normalized amplitude *absolute values* for the transmitted components $|\vec{E}_{0t}|_\perp$ and $|\vec{E}_{0t}|_\parallel$ can be expressed as:

$$\frac{|\vec{E}_{0t}|_\perp}{|\vec{E}_{0i}|_\perp} = t_\perp = -\frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i+\theta_t)} \text{ and } \frac{|\vec{E}_{0t}|_\parallel}{|\vec{E}_{0i}|_\parallel} = t_\parallel = +\frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i+\theta_t)\cos(\theta_i-\theta_t)} \quad (6.7)$$

If the incident wave is approaching normal incidence perpendicular to the planar sensor interface θ_i, θ_r and $\theta_t \to 0$. This implies that $\sin(\theta_i \pm \theta_t) \approx \tan(\theta_i \pm \theta_t) \approx (\theta_i \pm \theta_t)$ and by employing Snell's law $n_i\theta_i = n_t\theta_t$, the following expressions for $r_\perp$, $r_\parallel$, $t_\perp$ and $t_\parallel$ can be derived:

$$\left(r_\parallel\right)_{\theta_i=0} = -\left(r_\perp\right)_{\theta_i=0} = \frac{n_t-n_i}{n_t+n_i} \quad \text{and} \quad \left(t_\parallel\right)_{\theta_i=0} = -\left(t_\perp\right)_{\theta_i=0} = \frac{2n_i}{n_t+n_i} \quad (6.8)$$

Substituting for an air ($n_i = 1$) to glass interface ($n_t = 1.5$), the normalized reflected amplitudes amount to ± 0.2 complementing the ± 0.8 for the normalized transmitted component.

6.2.2 Energy transport in reflectance and transmittance at the sensor interface

The flow of electromagnetic energy through free space, associated with a traveling EM-wave, is represented by the Poynting vector $\vec{S} = c^2\epsilon_0\vec{E}\,\mathbf{x}\,\vec{B}$. This vector product represents the direction and magnitude of the energy transport per unit time across a unit area whose normal is parallel to $\vec{S}$. The radiant flux density or irradiance I (Watt m^{-2}) for a monochromatic wave with constant amplitude vector $\vec{E}_0$ (see above) is the average over time of the magnitude of the Poynting vector:

$$I = \overline{|\vec{S}|} = \frac{c\epsilon_0}{2}|\vec{E}_0|^2 \text{ in vacuum (free space)} \quad (6.9)$$

and

$$I = \overline{|\vec{S}|} = \frac{v\boldsymbol{\epsilon}}{2}|\vec{E}_0|^2 \text{ in a linear, homogeneous, isotropic dielectric} \quad (6.10)$$

with c the speed of light in vacuum, $v = c/n$ ($n = \sqrt{\epsilon_r\mu_r}$) the phase velocity of light in the dielectric and $\boldsymbol{\epsilon} = \epsilon_0\epsilon_r$.
Let I_i, I_r and I_t be the incident, reflected and transmitted radiant flux densities (irradiances), respectively. The radiation power arriving under an angle θ_i over an interface surface area A then equals $I_iA\cos\theta_i$, the reflected power $I_rA\cos\theta_r$ en the transmitted power $I_tA\cos\theta_t$. The **reflectance** R is defined by the ratio of the reflected radiant flux density to the incident radiant flux density, likewise the **transmittance** T is defined as the ratio of the transmitted radiant flux density to the incident radiant flux density, i.e.:

$$R = \frac{I_rA\cos\theta_r}{I_iA\cos\theta_i} \quad \text{and} \quad T = \frac{I_tA\cos\theta_t}{I_iA\cos\theta_i} \quad (6.11)$$

The quotient I_r/I_i equals $(v_r\boldsymbol{\epsilon_r}|\vec{E}_{0r}|^2/2)/(v_i\boldsymbol{\epsilon_i}|\vec{E}_{0i}|^2/2)$, but since the incident and the reflected radiation is in the same medium $v_r = v_i$, $\boldsymbol{\epsilon_r} = \boldsymbol{\epsilon_i}$ and $\theta_r = \theta_i$, the expression

for R becomes:

$$R = \frac{|\vec{E}_{0r}|^2}{|\vec{E}_{0i}|^2} = r^2 \tag{6.12}$$

The quotient I_t/I_i equals $(v_t\boldsymbol{\epsilon_t}|\vec{E}_{0t}|^2/2)/(v_i\boldsymbol{\epsilon_i}|\vec{E}_{0i}|^2/2)$. Using the fact that the permeability $\boldsymbol{\mu}$ $(= \mu_0\mu_r)$ for most of the relevant materials (insulators) have μ_r-values very close to one, it is therefore justified to assume $\boldsymbol{\mu_i} = \boldsymbol{\mu_t} = \mu_0$. This leads to the equalities $v_t\boldsymbol{\epsilon_t} = c\epsilon_0 n_t$ and $v_i\boldsymbol{\epsilon_i} = c\epsilon_0 n_i$, yielding the following expression for T:

$$T = \left(\frac{n_t \cos\theta_t}{n_i \cos\theta_i}\right)\frac{|\vec{E}_{0t}|^2}{|\vec{E}_{0i}|^2} = \left(\frac{n_t \cos\theta_t}{n_i \cos\theta_i}\right)t^2 \tag{6.13}$$

It is obvious that T cannot simply be equal to t^2: there is a difference in speed in the transportation of energy, expressed by the ratio of refractive indices and, secondly, the cross-sectional areas of the incident and transmitted beams are different which accounts for the ratio of the cosine terms.

Conservation of energy requires:

$$R + T = 1, \quad \text{also} \quad R_\perp + T_\perp = 1 \quad \text{and} \quad R_\parallel + T_\parallel = 1 \tag{6.14}$$

including the component values for R and T:

$$R_\perp = r_\perp^2 \quad R_\parallel = r_\parallel^2 \qquad T_\perp = \left(\frac{n_t \cos\theta_t}{n_i \cos\theta_i}\right)t_\perp^2 \quad T_\parallel = \left(\frac{n_t \cos\theta_t}{n_i \cos\theta_i}\right)t_\parallel^2 \tag{6.15}$$

These reflectance and transmittance components for an air to glass interface are illustrated as a function of incidence angle in Figure 6.2.

When the incident beam is perpendicular to the interface plane ($\theta_i = 0$), the incidence plane becomes undefined and all distinction between the parallel and perpendicular components vanishes. In this case the following important expressions for the distribution of radiant power at the sensor interface hold:

$$R = R_\perp = R_\parallel = \left(\frac{n_t - n_i}{n_t + n_i}\right)^2 \qquad T = T_\perp = T_\parallel = \frac{4n_i n_t}{(n_t + n_i)^2} \tag{6.16}$$

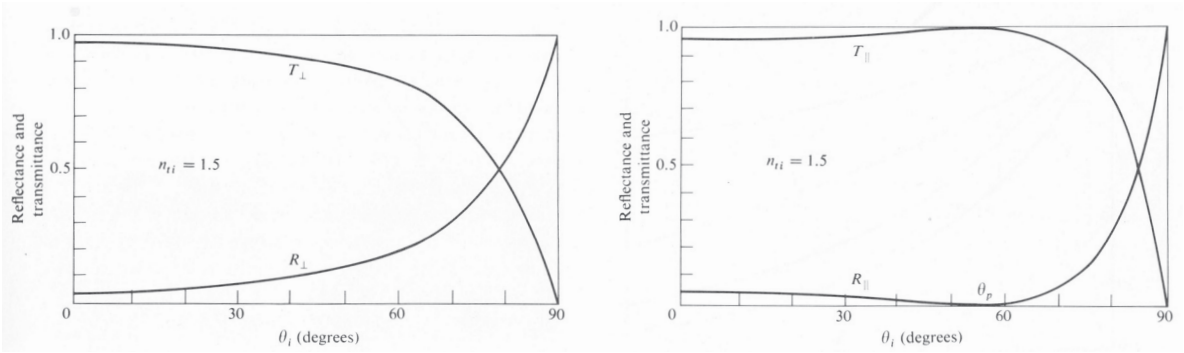


Figure 6.2: *Reflectance and Transmittance for an air to glass interface as a function of incidence angle θ_i, credit Hecht(1987).*

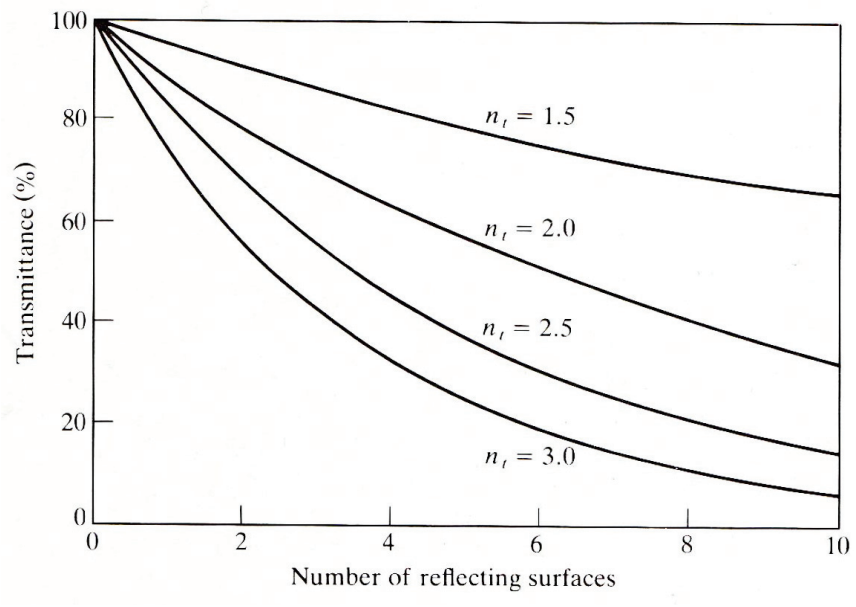


Figure 6.3: *Transmittance as a function of the number of slides in an optical stack for various values of the refractive index, credit Hecht(1987).*

From the expression for R it can be seen that 4% of the light that is incident normally on an air to glass interface will be reflected back. Figure 6.3 shows the deterioration of the transmittance with an increasing number (up to 10) of slides comprising non-absorbing materials with different refractive indices, a stack of 10 glass slides reduces the irradiance to two-third of its incident value. This indicates that working with complex lens systems introduces a substantial loss in light throughput, owing to the potentially large number of air-glass boundaries, say 10 to 20. Actually, when looking perpendicular into a stack of some 50 cover glass slides, about 87% of the incidence light will be reflected [transmission $(0.96)^{50}$] and the stack then acts very much like a real mirror. This is why a roll of 'clear' plastic tape is not transparent and why complex lens systems require anti-reflection coatings on many of the individual optical elements.

6.3 Detection of a radiation field by the sensor medium

Detection of a radiation field can generally be subdivided in two categories: amplitude detection and power detection.

6.3.1 Amplitude detection

This type of detection involves the measurement of the instantaneous amplitude of an electromagnetic wave with average frequency $\bar{\nu}$ and varying amplitude and phase.

Considering one component of polarization, the signal $S(t)$ varies as

$$S(t) = E_0(t)\cos(2\pi\bar{\nu}t + \phi(t)) \tag{6.17}$$

and is obtained by conversion of the electromagnetic field into a current by a receiving antenna. The relation between the signal and the amplitude of the wave can be regarded as linear, moreover the phase information $\phi(t)$ remains available. This detection process is therefore referred to as *coherent detection*. We shall treat this type of detection in the next chapter with the specific case of *superheterodyne coherent detection* by a quadratic sensing element of radio waves, micro waves and far infrared waves.

6.3.2 Power (or intensity) detection

In this case the detection process involves the measurement of the average power of an electromagnetic wave, a particle or a neutrino beam, in which the averaging takes place over a certain time interval ΔT, i.e. the detection yields a running average over ΔT. For electromagnetic waves this integration interval ΔT is large compared to the period of the wave $c/\bar{\nu}$, the signal $S(t)$ follows from:

$$S(t) = \frac{1}{\Delta T}\int_{t}^{t+\Delta T} \tilde{E}(t')\cdot\tilde{E}^*(t')\,dt' = \frac{1}{\Delta T}\,\Pi\left(\frac{t}{\Delta T}\right) * |\,\tilde{E}(t)\,|^2 \tag{6.18}$$

If the electromagnetic wave is regarded as a photon beam, the equivalent expression is:

$$S(t) = \frac{1}{\Delta T}\int_{t}^{t+\Delta T} n(t')\,dt' \tag{6.19}$$

in which $n(t')$ represents the photon flux. This relation also applies to corpuscular radiation (cosmic rays) and neutrinos. As discussed in Chapter 3, $n(t')$ relates to a Poisson process. If the time interval ΔT is taken sufficiently small, equation 6.19 reduces to single photon (particle) counting.

In this course no detailed treatment of the mechanism of amplitude detection will be given, however in the following paragraphs the physical principles of photon detection will be briefly reviewed, since they underlay the working principles of a large variety of radiation sensors.

6.4 Absorption of electromagnetic radiation

6.4.1 Opacity and absorption cross-sections

Consider a beam of electromagnetic radiation with monochromatic intensity $I(\epsilon, 0)$ at normal incidence on a layer of absorbing material (e.g. a detection element). The absorption in an infinitesimal layer ds within the absorber at depth s can be derived from the relation:

$$I(\epsilon, s+ds) - I(\epsilon, s) = -dI(\epsilon, s) = I(\epsilon, s)\,k(\epsilon, s)\,ds \tag{6.20}$$

The parameter $k(\epsilon, s)$ represents the probability per unit path-length that a photon is either absorbed or scattered. In general $k(\epsilon, s)$ is a function of the position s, but for a uniform homogeneous absorber $k(\epsilon, s) = k(\epsilon)$, independent of s. $k(\epsilon, s)$ is commonly referred to as the (volume) opacity or total absorption (extinction) coefficient of the relevant medium. Also $k(\epsilon, s) = [l(\epsilon, s)]^{-1}$, in which $l(\epsilon, s)$ represents the mean free path of a photon with energy ϵ. In practice a related quantity

$$\kappa(\epsilon, s) = k(\epsilon, s)/\rho \tag{6.21}$$

is often used, where ρ is the specific mass. $\kappa(\epsilon, s)$ is called the mass absorption coefficient in units of surface area per unit mass (e.g. $\mathrm{m^2 kg^{-1}}$). The absorption coefficient per atom for a uniform absorber follows from:

$$\sigma(\epsilon) = \frac{\kappa(\epsilon)\ A_m}{N_A} \tag{6.22}$$

where A_m represents the atomic mass and N_A Avogadro's number. The quantity $\sigma(\epsilon)$ (or $\sigma(\lambda)$, or $\sigma(\nu)$) is called the absorption (scattering) cross-section and needs to be used when describing the microscopic properties of an absorbing medium. The volume opacities $k(\epsilon), k(\lambda), k(\nu)$ are usually applied when treating radiation transport phenomena on a macroscopic level, such as stellar atmospheres.

Integration of equation 6.20, taking a total absorption length s_0, yields the attenuated beam intensity $I(\epsilon, s_0)$ after passing through s_0:

$$I(\epsilon, s_0) = I(\epsilon, 0)\ \exp\left(-\int_0^{s_0} k(\epsilon, s)\ ds\right) = I(\epsilon, 0)\ \exp(-\tau(\epsilon, s_0)) \tag{6.23}$$

in which

$$\tau(\epsilon, s) = \int_0^{s} k(\epsilon, s')\ ds' \tag{6.24}$$

represents the *optical depth*.

In order to understand the absorption potential and properties of various materials, it is necessary to treat the nature of the physical interaction processes which give rise to this absorption effect. In addition, the associated interaction products need to be assessed, since they comprise the conversion products which make up the physical signal at the detector output.

Basically the detection mechanisms for electromagnetic radiation can be divided into two major subclasses: the class in which during the interaction process *no free charge carriers* are produced and the class where the interaction *does lead to the production of free charge carriers*.

The charge carriers in the target material of the sensing element that are involved in these interaction processes can be subdivided in four categories:

- Electrons located in the innermost shells of the constituent atoms. The interaction cross-section solely depends on the properties of the individual atoms, excitation or ionization demands photons with a relatively high energy $\epsilon = \hbar\omega$, like UV, X-ray or

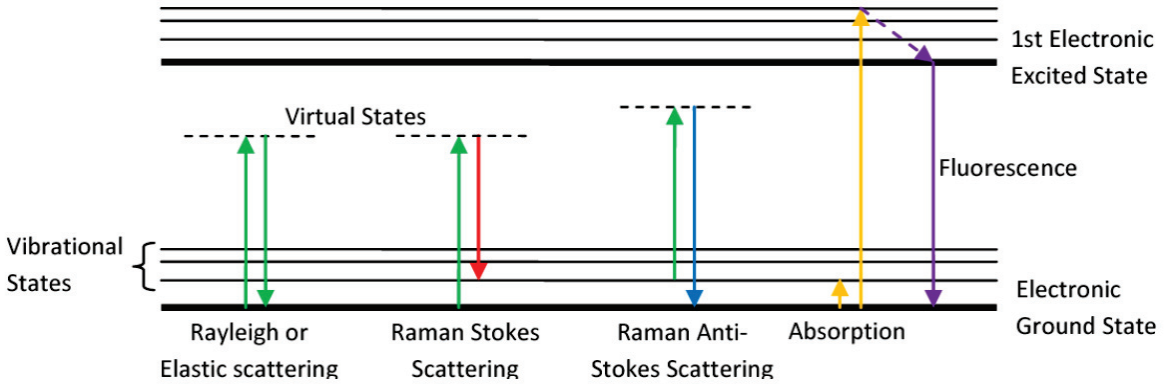


Figure 6.4: *Energy transitions in elastic (Rayleigh) and inelastic (Raman-Stokes) photon scattering processes.The phonon production resulting from the inelastic scattering process raises the mechanical energy contained in the lattice vibrations.*

γ-ray photons. The magnitude of the absorption coefficient $k(\epsilon, s)$is strongly dependent on the nuclear charge Z of the atomic nucleus, the cross-section for photo-electric ionization is proportional to Z^5, except for photon energies close to the absorption edges that mark the binding energies of the orbital electrons.
- Electrons in the valence band.
- Electrons bound to local impurities, like donor atoms or crystal imperfections.
- Free charge carriers in the conduction band.

6.4.2 Interaction processes

Interaction processes that do not release free charge carriers

- Inelastic scattering of a photon in a molecular electron cloud (Raman-Stokes scattering), the energetics of the process is shown in figure (6.4). In contrast to Rayleigh and Mie scattering, where the photon only changes direction but loses no energy (elastic scattering), the photon transfers mechanical energy to the crystal lattice that excites to a higher vibrational state through a quantized energy step that is characterized by a phonon: a package of quantized thermal energy. Lattice vibrations at a certain temperature T produce a traveling *elastic* wave with wave vector $\vec{s}$ (phonon momentum $\vec{p} = \hbar\vec{s}$) and frequency ν_s, so we have:

$$h\nu_{rad} = h\nu'_{rad} + h\nu_s, \tag{6.25}$$

  with ν_{rad} the frequency of the incident photon, ν'_{rad} the frequency of the scattered photon and $h\nu_s = \hbar a_s|\vec{s}|$ the phonon energy (a_s = sound velocity in the sensor medium). This process results in a red shift of the incident photons, an effect that is employed in wavelength shifters.

- 'Exciton' generation. Another process that falls in this category concerns the production of a so-called 'exciton'. This constitutes the production of a coupled electron-hole pair that energetically resides in the forbidden zone, these electron-hole combinations move as pairs in the crystal lattice. The existence of these

exciton energy levels has been first investigated in the photon absorption spectra of alkali-halides (metal-halogen salts comprising atomic bonds between Na, K, Cs and Cl, Br, I). These absorption spectra showed distinct peaks at certain wavelengths that could not be associated with enhanced photo-conduction, indicating that no extra free charge carriers were produced. The explanation for this phenomenon is thought to be that the halogen ions become excited to a small band of energy levels that resides below the conduction band in the forbidden energy gap, the so-called exciton bands. The fact that the electron-hole pair moves as a unity is explained by the assumption that they stay locked in each others Coulomb field. In semiconductors, as a consequence of morphological structures like surface irregularities and impurities, exciton energy levels in the forbidden gap may be permanently present that are *partly free* for electron occupation following excitations from the valence band.

- Complete photon ⇒ phonon conversion. Photon interaction through various processes, i.e. Raman-Stokes and Brillouin scattering, excitation and relaxation in vibrational and rotational states by free electron scattering, can result in a complete conversion of the photon energy into phonons. This energy transfer process can be regarded as a transformation of the incident radiant energy into mechanical energy of the lattice vibrations (i.e. in 'heat') in the target material, so we have:

$$\epsilon = h\nu_{rad} \quad \Rightarrow \quad \sum_i (h\nu_s)_i \tag{6.26}$$

in which $h\nu_s$ represents the characteristic phonon energy ($\approx kT_{material}$).
Thermal devices respond to the "heating" effect of the incident radiation field by changing their temperature. This process requires two steps: the radiation field changes the temperature of the absorber and subsequently this temperature change causes or induces some measurable parameter change:

Device:		*Physical parameter change*
Bolometer	⇒	Resistance
Thermocouple	⇒	Voltage
Pyroelectric	⇒	Capacitance in ferro-electric material
Super-conductor	⇒	Resistance

The absorption characteristic of a thermal device is normally expressed by the so-called absorptance $\alpha(\lambda)$. *Kirchhoff's law* for electromagnetic radiation states that in thermodynamic equilibrium the amount of energy absorbed is equal to the amount of energy emitted. Therefore, for all surfaces:

$$\alpha(\lambda) = \mathcal{E}(\lambda) \tag{6.27}$$

with $\mathcal{E}(\lambda)$ the spectral emissivity and $\alpha(\lambda)$ the ratio of the amount of radiation absorbed to the amount of radiation incident monochromatically:

$$\alpha(\lambda) = \frac{\{I(\lambda)\}_{abs}}{\{I(\lambda)\}_{incid}} \tag{6.28}$$

A blackbody has a spectral emissivity $\mathcal{E}(\lambda) = 1$ and therefore absorbs all incident radiation, it can reflect no light and, hence, appears black. Thermal sensors are devices where the energy absorbed depends often on the *surface properties* of the material involved. The spectral response is determined by the spectral dependence of the surface absorptance:

$$\alpha(\lambda, T) = \mathcal{E}(\lambda, T) = \frac{I(\lambda, T)_{material}}{I(\lambda, T)_{blackbody}} \tag{6.29}$$

Values of "thermal mass" (= specific heat × mass) and thermal conductance are basic parameters for a thermal device and determine its response time. Thermal detectors are typically slower responding devices than photo-electric sensors since their thermal mass must experience a rise and fall of temperature. A large temperature change per unit radiation power requires a small thermal mass or a well-isolated detector. Consequently the heat injection due to the radiation field will not dissipate very fast and the decay time becomes relatively long.
In addition the spectral response for a thermal device is much broader than a typical photo-electric sensor and extends into the far-infrared (up to several hundred micron).

Interaction processes that do produce free charge carriers

Photo-electric devices detect radiation by direct interaction of the individual photons with the atomic structure of the material (or with the free atoms in a gas). This can be quantitatively described with the aid of the linear absorption coefficient $k(\epsilon, \lambda, \nu)$ or the cross-section $\sigma(\epsilon, \lambda, \nu)$ as discussed in the previous section. This interaction of radiation with matter produces parameter changes that can be detected by associated circuitry or interfaces. Physical parameters that can change in these devices due to irradiation comprise resistance, inductance (due to charge generation), voltage and current.

- The external photo-electric effect (photo-emission effect).
  The effect is based on the release of a free electron by an incident photon, moreover in solids a phonon is also released. Therefore the reaction is given by:

  $$h\nu \Rightarrow \text{photo-electron}(+h\nu_s) \tag{6.30}$$

  The photo-electron physically leaves the target material and becomes independent of the properties of the material from which it was emitted. We have:
  In a gas (free atoms/molecules):

  $$h\nu = W + \frac{1}{2} m_e v^2 \Rightarrow (\text{electron ionization energy + kinetic energy}) \tag{6.31}$$

  In a metal (lattice structure):

  $$h\nu > qU_M,\ U_M = \Phi_V - E_F/q = \text{metal workfunction} \tag{6.32}$$

  For most metals qU_M has a value of $3 \Rightarrow 5$ eV. With $(h\nu - qU_M)$ becoming increasingly positive, the effective cross-section for this photo-emission process increases accordingly.There can be a significant influence on the yield due to

irregularities of the surface structure of the electron-emitting material. This will result in a lack of uniformity in the photo-emission process of the material under consideration. This process is widely applied in the use of photo-cathodes in photomultiplier tubes and micro-channel plates for low light-level detection and night-viewing.

- The internal photo-electric effect.
  As the name indicates, surface effects cannot play a role in this case, this photo-electric effect constitutes an internal process within the target material. We can distinguish two different processes:
  1. The process takes place between two dissimilar materials that are electrically coupled like a pn-junction or a Schottky contact. If the energy transfer between the photon and the electron is sufficiently large for the electron to overcome the internal potential barrier of the junction, this will result in a photon driven current through the contact (photo-current). Obviously the energy of the photon needs to be larger than the contact potential, so for a pn-junction we have $h\nu \geq q\Phi_{pn}$.
  2. The process comprises the production of an electron-hole pair by photon-excitation of a valence electron to the conduction band. This comes on top of the thermally generated electron-hole pairs. In fact the latter can also be considered as an internal excitation process, i.e. the interaction between a valence electron and a phonon. The fraction of photons that actually do produce an electron-hole pair is designated as the *quantum efficiency*. The energetics required for this interaction would *seem to be*:

$$h\nu \geq E_c - E_v = E_{gap} \tag{6.33}$$

  However in practice this boundary turns out not to be sharp for a number of semiconductor materials, among which Ge and Si, since one has to discriminate between two types of transition:
  -A direct transition
  -An indirect transition
  Let us consider this in some detail. The photon excitation process is subject to the conservation laws of energy and momentum. The momentum of the electron $\vec{p}$ equals $\hbar\vec{k}$, where $\vec{k}$ represents the wave vector of the electron due to its wave character, furthermore the electron energy can be expressed as a function of $\vec{k}$, i.e. as $E(\vec{k})$. The incident photon can be characterized by a wave vector $\vec{\sigma}$ and energy $h\nu \approx E_e$. So we can write the following balance conditions:

$$E(\vec{k}') = E(\vec{k}) + h\nu \quad \text{(energy balance)} \tag{6.34}$$

$$\vec{k}' = \vec{k} + \vec{\sigma} \quad \text{(momentum balance)} \tag{6.35}$$

  The magnitude of the photon momentum $|\hbar\vec{\sigma}|$ in the infrared to ultraviolet wavelength range is negligible as compared to the the electron momentum $|\hbar\vec{k}| = \sqrt{2mE_e}$, since $|\vec{\sigma}|/|\vec{k}| = \sqrt{E_e/2mc^2} \approx 10^{-3}$ with $E_e = 1 - 4$ eV and $2mc^2 \approx 1$ MeV. The momentum balance for these *optical transitions* therefore dictates with $|\vec{\sigma}| \approx 0 \Rightarrow \vec{k}' \approx \vec{k}$, implying that optical transitions in the $E - \vec{k}$ diagram occur

vertically.
The simplest $E-\vec{k}$ diagram, i.e. for the case of free electrons, follows from the solution of the time independent Schrödinger equation with the potential term V set to zero:

$$\Delta\psi + \frac{2mE}{\hbar^2}\psi = 0 \tag{6.36}$$

If we consider a cube with edge L and apply the boundary conditions $\psi = 0$ for $x, y, z = 0$ and L respectively, this constrains the values of E to:

$$E = \frac{\hbar^2\pi^2}{2mL^2}(n_1^2 + n_2^2 + n_3^2) \tag{6.37}$$

where n_1, n_2, n_3 are integer quantum numbers ($n_i = 1, 2, 3, ...$), each energy quantum state is represented by a single point in n-space. Since these points are situated very close together, we might regard the n-space, and therefore also E, as a continuum. Substituting the wave vector $\vec{k}$ ($|\vec{k}| = 2\pi/\lambda$), with components k_x, k_y, k_z:

$$k_x = \frac{\pi}{L}n_1,\ k_y = \frac{\pi}{L}n_2,\ k_z = \frac{\pi}{L}n_3,\ \text{with } L = \frac{n_i\lambda}{2} \quad \Rightarrow \quad E = \frac{\hbar^2}{2m}|\vec{k}|^2 \tag{6.38}$$

So in this case $E(\vec{k})$ is a parabolic function.
In the solid state, however, the electrons are subject to a *periodic potential* $V(x, y, z)$ that leads to much more complicated solutions that result in permitted and forbidden zones for the energy E in the $E-\vec{k}$ diagram. In the one-dimensional Bloch model, i.e. an array of atoms at mutual distance d in a lattice, discontinuities ΔE occur in the $E-(\vec{k} = k_x)$ diagram for k_x-values equal to $n\pi/d$, the so-called Brillouin zones. A general treatment of the $E-\vec{k}$ relationship for the three-dimensional case shows that the energy minimum in the conduction band and the energy maximum in the valence band differ in the momentum value of $\vec{k}$. In that case a vertical transition in the $E-\vec{k}$ diagram ($\vec{k} = \vec{k}'$) requires a photon energy $h\nu > E_{gap}$.
Nevertheless, since $h\nu = E_{gap}$ transitions are observed, the explanation for these needs to be found in the occurrence of *indirect transitions* that involve the emission or absorption of a phonon (frequency ν_s, momentum $\vec{s}$). The optical transitions that occur are depicted in figure (6.5). The energy and momentum conservation laws for these indirect optical transitions can be written as:

$$E(\vec{k}') = E(\vec{k}) + h\nu \pm h\nu_s \approx E(\vec{k}) + h\nu, \text{ since } h\nu \gg h\nu_s \tag{6.39}$$

$$\vec{k}' = \vec{k} + \vec{\sigma} \pm \vec{s} \approx \vec{k} \pm \vec{s}, \text{ since } \vec{\sigma} \ll \vec{s} \tag{6.40}$$

At 300 K, the phonon energy equals ≈ 25 meV and can be regarded as almost negligible compared to the optical photon energies (1−4 eV). In contrast to this the photon momentum $h\nu/c$ is nearly negligible relative to the phonon momentum (division by the light velocity c compared to division by the sound velocity v_s).This means that the phonon vectors (green in figure 6.5) run almost horizontal in the E-$\vec{k}$ diagram. For a transition $h\nu = E_{gap}$ the phonon needs to have

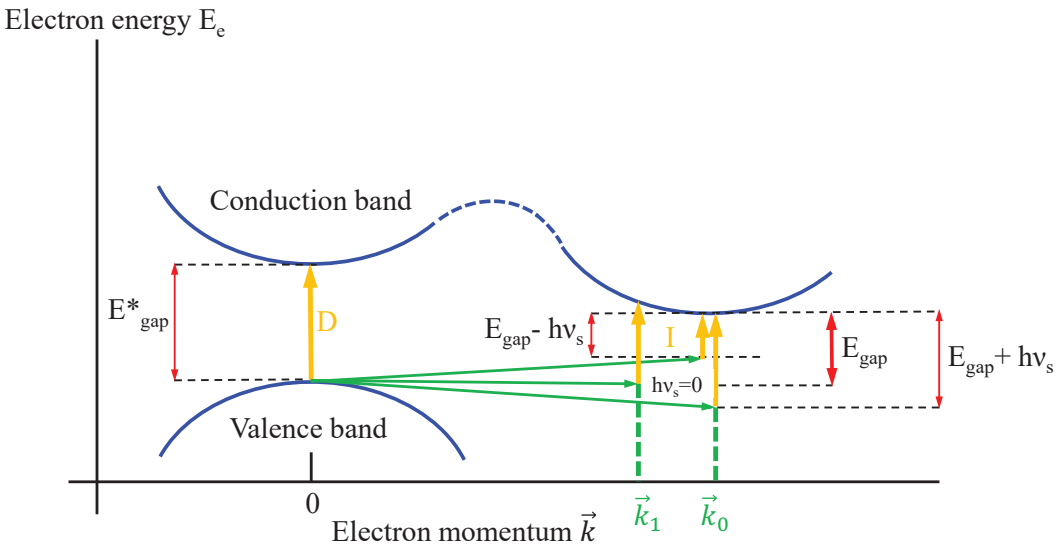


Figure 6.5: *Electron energy versus electron momentum diagram. Yellow vectors: direct (D) and indirect (I) optical transitions, green vectors: phonon (energy $h\nu_s$) contributions. E_{gap} minimum energy band gap for indirect transitions, E^*_{gap} energy band gap for direct transitions ($> E_{gap}$).*

a momentum of exactly $\hbar\vec{s} = \hbar\vec{k}_0$, which has low probability, however for somewhat higher photon energies $h\nu < E^*_{gap}$ a larger spectrum of phonon momenta ($\hbar\vec{s} = \hbar\vec{k}_1 \Rightarrow \hbar\vec{k}_0$, see figure 6.5) suffices to allow a photo-electric transition, which considerably enhances the transition probability.

The wavelength associated with an energy gap E_{gap} is $\lambda = 1.24/E_{gap}$, with E_{gap} expressed in eV and λ in μm. Figure (6.6) shows the absorption coefficients as a function of wavelength for four different semiconductor materials, two compounds (GaAs and InP) and two elements (Ge and Si). The compound semiconductors show a steep rise of the absorption, indicative for the dominance of direct transitions: for GaAs and InP the minimum energy of the conduction band and the maximum energy of the valence band occur at the same value of the momentum vector $\vec{k}$. The direct transition for GaAs at 300K is at a wavelength of 870 nm (1.424 eV), whereas InP (300 K) kicks in at 923 nm (1.344 eV), see figure(6.7).

In the case of Ge and Si, indirect transitions play a major role in the absorption characteristics, the absorption coefficient does not exhibit a steep slope as a function of wavelength but rather shows a gradual increase/decrease. Crystalline Si, with an indirect band gap at 1107 nm (1.12 eV), only shows direct transitions in the ultraviolet starting from 365 nm (3.4 eV, see figure 6.8). Ge exhibits an indirect band gap at 1878 nm (0.66 eV) with a multitude of indirect transitions at shorter wavelengths, a direct transition ($E_{\Gamma 1}$) lies in the short-wave infrared

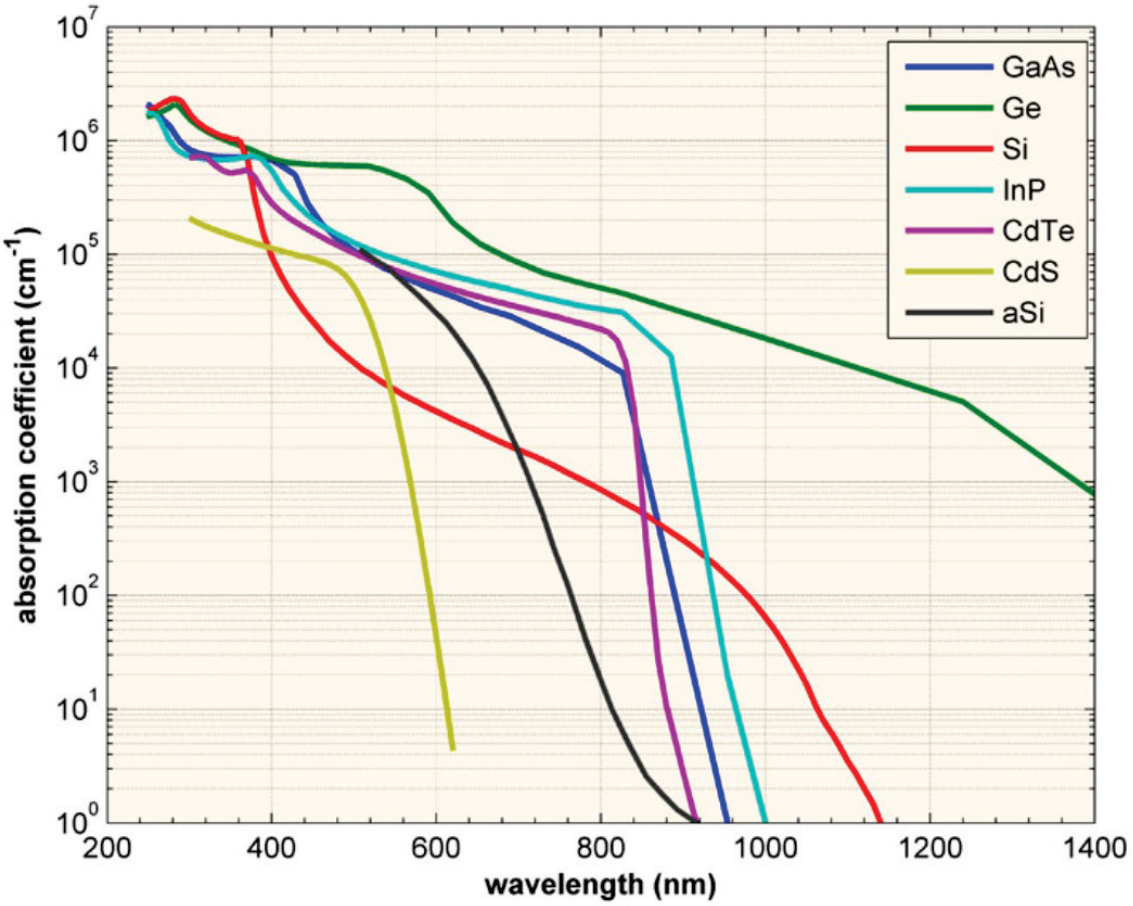


Figure 6.6: *Photo-electric absorption coefficients for seven semiconductor elements/compounds. GaAs, CdTe, CdS and InP are dominated by direct band gap transitions ('Direct band gap semiconductors'), Si and Ge are dominated by indirect band gap transitions ('Indirect band gap semiconductors'): Ge shows the longest wavelength response, CdS starts at the shortest wavelength.*

at 1550 nm (0.8 nm). Both Ge and crystalline Si are suitable for visible light detection, however the bulk wavelength sensitivity for Ge lies in the short-wave infrared band and it possesses reduced sensitivity as compared to Si when integrated over the visible light spectrum from 400−700 nm.
Detection of radiation in the near infrared with *intrinsic* semiconductors requires materials with $E_{gap} = 0.1 - 0.4$ eV ($\lambda = 3 - 13\mu m$). Suitable semiconductors for this purpose comprise:

InAs ($E_{gap} = 0.40$ eV at T=300 K)
InSb ($E_{gap} = 0.23$ eV at T=300 K)
HgCdTe ($E_{gap} = 0.09$ eV at T=300 K)
PbSnTe ($E_{gap} = 0.10$ eV at T=300 K)
The value of E_{gap} for the latter two compounds is tunable by choosing the molar fraction of Cadmium(Tin). The available range of E_{gap} and cut-off wavelength

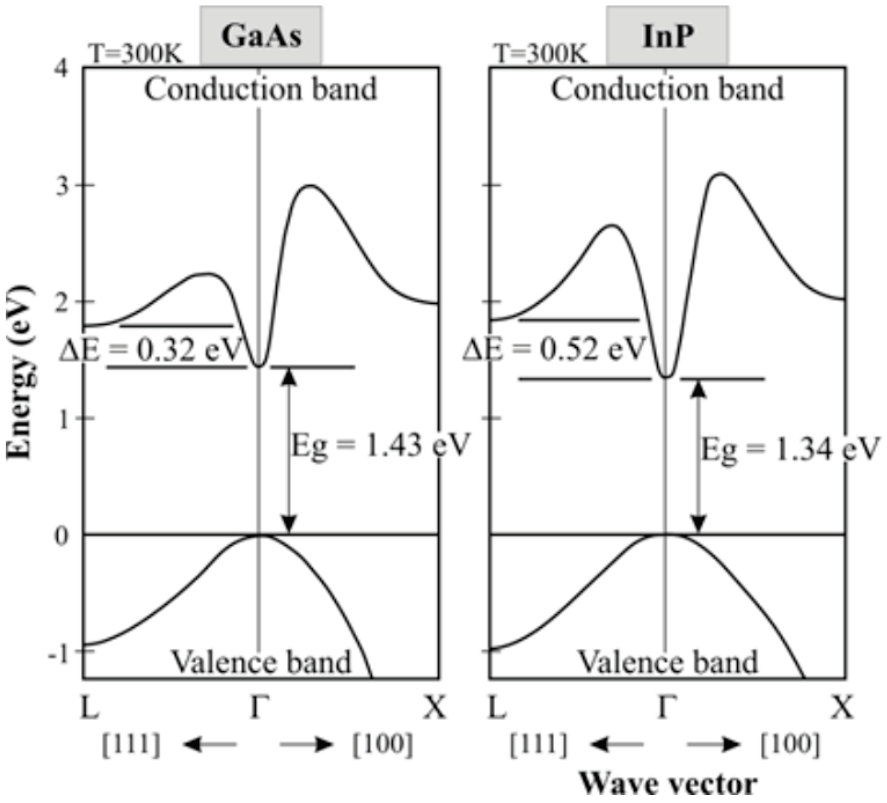


Figure 6.7: *$E - \vec{k}$ diagram of the direct band materials GaAs and InP, the band gap of the dominant direct photo-electric transition is shown: $\vec{k}$ coincides for minimum (maximum) energy value of the conduction (valence) band.*

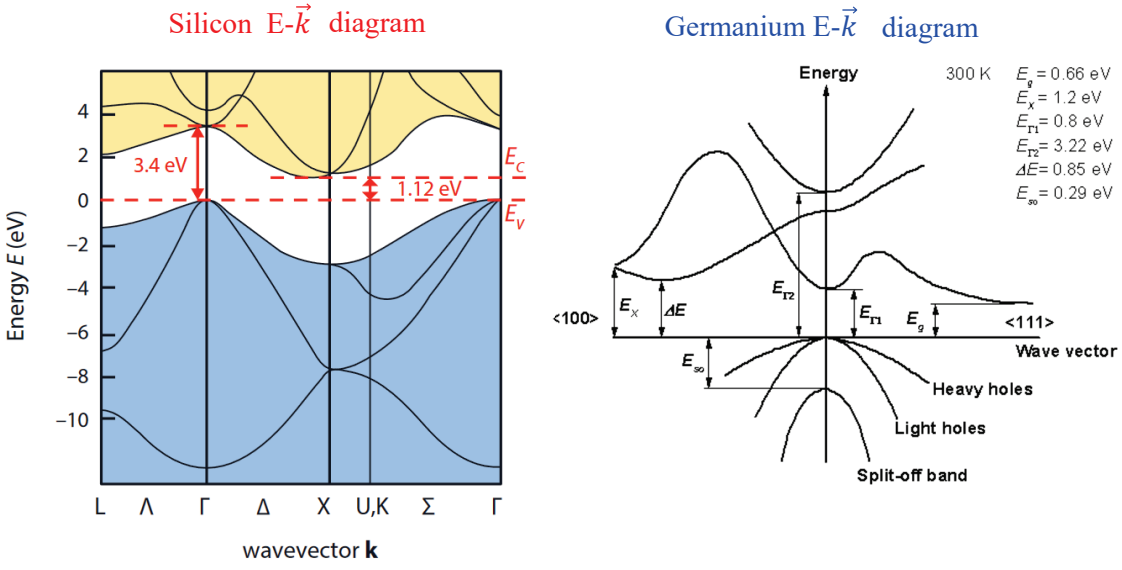


Figure 6.8: *Crystalline Silicon and Germanium: Band diagrams showing the positions of indirect and direct transitions in the $E - \vec{k}$ diagram.*

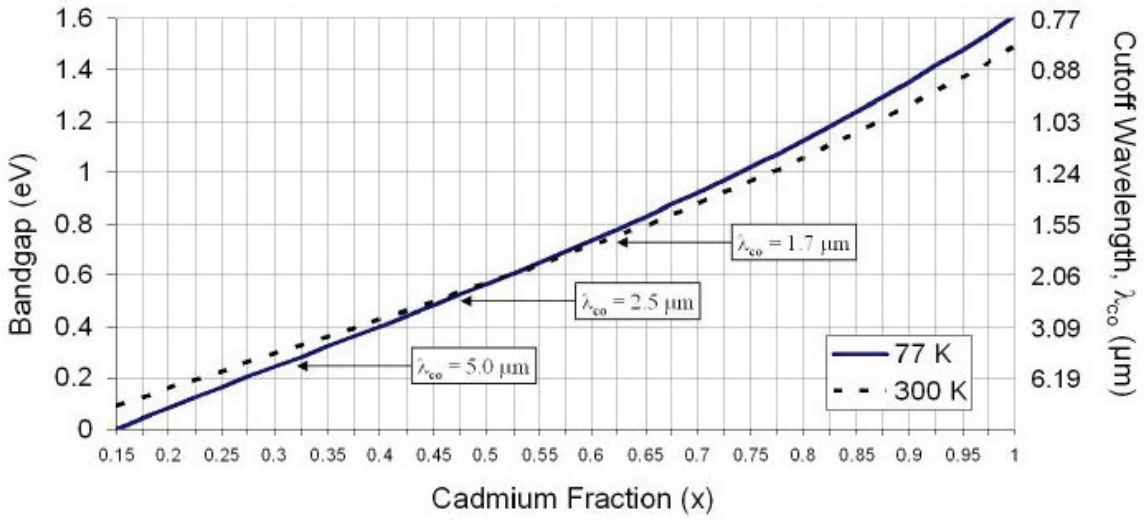


Figure 6.9: *Tunable band gap in the compound semiconductor HgCdTe through adjustment of the molar fraction of Cd(x) relative to Hg(1-x):* $Hg_{1-x}Cd_xTe$. *Fits with a polynomial in* x *and a temperature correction factor:* $E_{gap} = -0.302 + 1.93x - 0.81x^2 + 0.832x^3 + 5.35\,10^{-4}T(1-2x)$.

λ_{co} values for HgCdTe are shown in figure (6.9).
Extrinsic semiconductors can be employed for detection of far-infrared radiation (100 μm+) by a suitable choice of donor or acceptor impurities. Obviously this will only be possible if the ambient is at cryogenic temperatures, since otherwise all the donor/acceptor levels will already have been ionized by thermal exchange with the sensor environment.

6.4.3 Figures of merit

A basic figure of merit that applies to all sensors with an electrical output is *responsivity*. By definition the responsivity comprises the ratio of the electrical output, usually in Volts or Amperes, to the radiant power input (i.e. Watts) onto the active area of the sensor. Hence for the *spectral voltage responsivity* $R_V(\lambda,\nu)$ and the *spectral current responsivity* $R_I(\lambda,\nu)$ of a sensor irradiated by a source of wavelength λ that is modulated at a frequency ν we have:

$$R_V(\lambda,\nu) = \frac{V_s}{\phi(\lambda)}\ \text{[Volt/Watt]} \qquad \text{and} \qquad R_I(\lambda,\nu) = \frac{I_s}{\phi(\lambda)}\ \text{[Ampere/Watt]}, \quad (6.41)$$

in which $\phi(\lambda)$ represents the spectral radiant power incident on the active area of the sensor.
Alternatively, the *blackbody* responsivities $R_V(T,\nu)$ and $R_I(T,\nu)$ are the the voltage and current sensor outputs divided by the incident radiant power from a black body

source of temperature T modulated at a frequency ν that produces those outputs:

$$R_V(T,\nu) = \frac{V_s}{\int_0^\infty \phi_{bb}(\lambda)d\lambda} \qquad \text{and} \qquad R_I(T,\nu) = \frac{I_s}{\int_0^\infty \phi_{bb}(\lambda)d\lambda}, \tag{6.42}$$

where $\phi_{bb}(\lambda)$ now represents the blackbody (Planck) spectral distribution. If we consider a black body of temperature T with an area A_{bb} irradiating a sensor with active area A_s at a distance R (implicitly both areas normal to the optical axis), the blackbody voltage responsivity equals $\pi R^2 V_s/(A_{bb}A_s\sigma_{SB}T^4)$, with σ_{SB} the Stefan Boltzmann constant.
Employing the responsivity we can define another figure of merit that is widely in use: the *Noise Equivalent Power* (NEP). This NEP is the required power incident on a sensor system to produce a signal output that is equal to the rms noise output. Phrased differently, the NEP is the signal level that produces a signal-to-noise ratio of 1. Utilizing the current responsivity, we obtain two alternative expressions for the NEP:

$$\begin{aligned} I_{rms} &= R_I(\lambda,\Delta\nu_N)NEP(\lambda,\Delta\nu_N) \Rightarrow \\ NEP(\lambda,\Delta\nu_N) &= \frac{I_{rms}}{R_I(\lambda,\Delta\nu_N)}\left(=\frac{V_{rms}}{R_V(\lambda,\Delta\nu_N)}\right) \quad \text{[Watt]} \end{aligned} \tag{6.43}$$

$$\begin{aligned} I_{rms} &= R_I(T,\Delta\nu_N)NEP(T,\Delta\nu_N) \Rightarrow \\ NEP(T,\Delta\nu_N) &= \frac{I_{rms}}{R_I(T,\Delta\nu_N)}\left(=\frac{V_{rms}}{R_V(T,\Delta\nu_N)}\right) \quad \text{[Watt]} \end{aligned} \tag{6.44}$$

with I_{rms} (V_{rms}) the root-mean-square noise current (voltage) within the equivalent bandwidth of the noise signal $\Delta\nu_N$. $R_I(\lambda(T),\Delta\nu_N)$ and $R_V(\lambda(T),\Delta\nu_N)$ represent the *average* responsivities over the noise bandwidth $\Delta\nu_N$.
Intermezzo:
The sensor elements treated in these lectures all show an intrinsic high-frequency cut-off corresponding to the frequency response of a first order system with a characteristic time constant τ_c. The responsivity of a first order system can be expressed as:

$$R(\lambda,\nu) = \frac{R(\lambda,0)}{\sqrt{(1+4\pi^2\nu^2\tau_c^2)}} \tag{6.45}$$

In describing the noise characteristics with the power spectral density $S(\nu)$ we have defined (see expression (5.83)) a one-sided noise equivalent bandwidth (NEBW):

$$\Delta\nu_c = \int_0^\infty \frac{d\nu}{(1+4\pi^2\nu^2\tau_c^2)} = \frac{1}{4\tau_c} \tag{6.46}$$

which is slightly larger than the roll-off frequency tipping point $\nu_{ro} = 1/(2\pi\tau_c)$ in the bode diagram for a first order system. Formally we then need to compute the *average responsivity* $\overline{R(\lambda,\nu)}$ over *the noise equivalent bandwidth*:

$$\overline{R(\lambda,\nu)} = 4\tau_c R(\lambda,0)\int_0^{1/(4\tau_c)} \frac{d\nu}{\sqrt{(1+4\pi^2\nu^2\tau_c^2)}} \Rightarrow \frac{2R(\lambda,0)}{\pi}\int_0^{\pi/2}\frac{dx}{\sqrt{1+x^2}}, \quad x = 2\pi\tau_c\nu \tag{6.47}$$

Solving the integral and substituting the boundary values we get:

$$\overline{R(\lambda,\nu)} = R(\lambda,\Delta\nu_c) = \frac{2R(\lambda,0)}{\pi}\ell n\left[\sqrt{1+\left(\frac{\pi}{2}\right)^2}+\frac{\pi}{2}\right] \approx 0.8R(\lambda,0) \tag{6.48}$$

In what follows we shall mostly neglect this numerical factor, since it is *of order 1*, and rather utilize $R(\lambda,0)$ formally only representative for the low frequency regime $\nu \ll 1/(2\pi\tau_c)$.
End intermezzo

The spectral NEP$(\lambda,\Delta\nu_N)$ is the monochromatic radiant flux required to produce an rms signal-to-noise ratio of 1 in a noise bandwidth $\Delta\nu_N$, whereas the blackbody NEP$(T,\Delta\nu_N)$ constitutes the blackbody radiant flux required to produce an rms signal-to-noise ratio of unity for a noise bandwidth $\Delta\nu_N$. A low NEP implies that a weak signal can be detected, hence the lower the NEP the more sensitive the sensor. The NEP is a useful quantity when comparing similar sensors under identical conditions, but it cannot be sensibly used for sensor performance comparison between dissimilar detectors. After all the NEP is a function of the parameters under which it is measured such as the active sensor area and the actual frequency bandwidth that directly determine its magnitude, however neither of these parameters were specified in the definition of the NEP. A comparison of noise equivalent powers measured under different conditions can therefore lead to misleading conclusions.
To compensate for the intrinsic dependence on sensor collecting area and specific frequency bandwidth of the actual values of the NEP, a more useful figure of merit is the

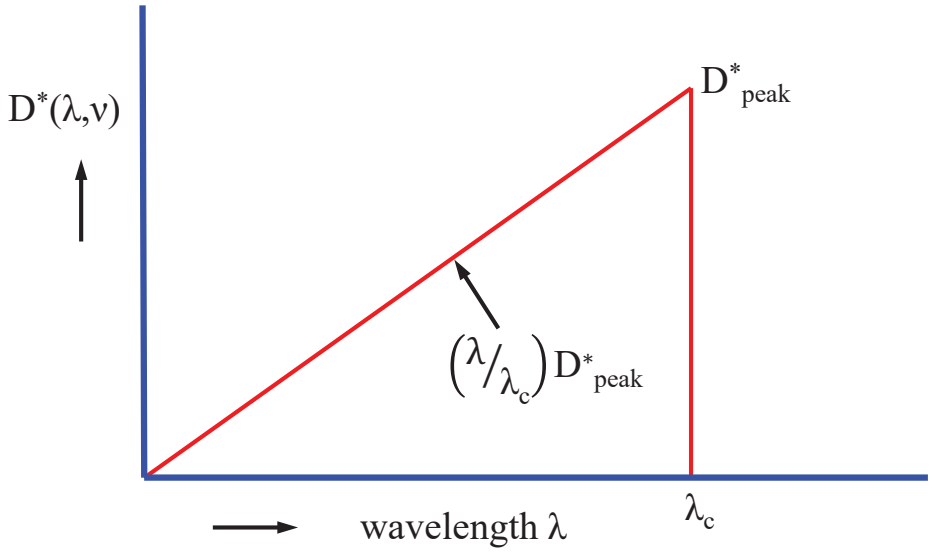


Figure 6.10: *The spectral normalized detectivity $D^*(\lambda,\nu)$ as a function of wavelength λ. The value of λ_c is determined by the threshold of the sensor material for photogeneration of free charge carriers, this also marks the position for the peak value of D^*. The slope towards shorter wavelengths is governed by the ratio of λ to λ_c.*

so-called normalized detectivity D*, defined as:

$$D^* = \frac{\sqrt{A_s \Delta\nu_N}}{NEP} = \frac{\sqrt{A_s \Delta\nu_N}}{\phi(\lambda)}(S/N) \qquad [\text{ cm Hz}^{1/2}\text{ Watt}^{-1}], \tag{6.49}$$

where either the spectral or the blackbody value of the NEP may be utilized to define $D^*(\lambda, \Delta\nu_N)$ or $D^*(T, \Delta\nu_N)$. The advantage of using D* as a figure of merit is its normalization to an unit detector area [e.g. cm^2] and to an unit frequency noise bandwidth of 1 Hertz. Another way of interpreting D* is that it is equal to the signal-to-noise ratio at the output of the sensor when 1 Watt of radiant power is incident on a sensor area of 1 cm^2 with a bandwidth of 1 Hz, see the far right-hand expression in (6.49). This interpretation is of course only meant as a mental concept, since a radiative power of 1 Watt is far off-scale compared to the actual values encountered in practical situations. Nevertheless, owing to its normalization, D* is a useful parameter to compare directly the merit of sensors of different size whose performance was measured using different bandwidths. Do recall however that the important assumptions made in arriving at a relevant usage of D* are that the noise is proportional to $\sqrt{\Delta\nu_N}$ with a flat (white) power spectral density distribution and that the noise is proportional to the sensor active area $\sqrt{A_s}$.
The wavelength dependence of the spectral normalized detectivity $D^*(\lambda, \Delta\nu_N)$ carries two characteristic features as a result of the quantum character of the detection process, i.e. the conversion of photons into charge carriers starting from the minimum photon energy $h\nu_c$ required to produce free charge carriers in the sensor material ($h\nu_c = hc/\lambda_c \Rightarrow$ maximum wavelength λ_c. Furthermore, any radiation incident on the sensor at wavelengths shorter than λ_c will have a $D^*(\lambda, \Delta\nu_N)$ reduced from the peak spectral D^* at λ_c proportional to the ratio λ/λ_c. The main spectral signatures of $D^*(\lambda, \Delta\nu_N)$ are shown in figure (6.10).

Chapter 7

Radiation sensors: sensor elements

7.1 'Quadratic' sensing elements for wave detection

7.1.1 Thermal radiation: superposition of wave packets

A beam of thermal radiation will ordinarily comprise a myriad of randomly overlapping wave groups or *wave packets*. They arise from two different types of quasi-monochromatic sources:
- Gaussian shaped spectral lines which emerge when several line broadening mechanisms contribute to the line formation (*viz.* the central limit theorem).
- Spectral lines with a Lorentz profile, characteristic for the natural line broadening associated with the intrinsic time spread of radiative atomic transitions that are governed by the uncertainty relation of Heisenberg $\sigma_\epsilon \sigma_t = \hbar/2$.

The characteristic length, τ_c, of these wave packets in the time domain follows from the Fourier transform of their spectral frequency distribution.The characteristic time τ_c is commonly referred to as the *coherence time*, it represents the typical time scale over which the phase of the EM-wave can be predicted with reasonable accuracy at a *given location in space*. For atomic transitions in the optical $\tau_c \simeq 10^{-9}$s. Figure (7.1) shows a quasi-monochromatic signal (one degree of polarization) comprising a random superposition of individual wave packets.
This wave signal fluctuates both in amplitude and in frequency, the latter characterized by a typical bandwidth $\Delta\nu$ around an average frequency $\bar{\nu}$. The frequency stability of such a quasi-monochromatic wave is defined by $\bar{\nu}/\Delta\nu$.
The linearly polarized signal displayed in figure (7.1) can mathematically be expressed by the real function:

$$E(t) = E_0(t)\cos(2\pi\bar{\nu}t + \phi(t)) \tag{7.1}$$

The amplitude $E_0(t)$ of the quasi-monochromatic wave is a *wide-sense stationary* Gaussian random time function of zero mean. Moreover the stochastic process is assumed to be *mean- and correlation-ergodic*, i.e. for an arbitrary real stochastic variable $X(t)$

its *expectation value* at time t, $\mathbf{E}\{X(t)\}$, can be interchanged with its time average:

$$\bar{X} \;=\; \lim_{T\to\infty}\frac{1}{T}\int_{-\frac{1}{2}T}^{+\frac{1}{2}T} X(t)\;dt = \mathbf{E}\{X(t)\} \tag{7.2}$$

Furthermore, the expectation value at time t of the autocorrelation $R_X(\tau)$ of $X(t)$ can be interchanged with its time average:

$$R_X(\tau) \;=\; \overline{X(t)\cdot X(t+\tau)} \;=\; \lim_{T\to\infty}\frac{1}{T}\int_{-\frac{1}{2}T}^{+\frac{1}{2}T} X(t)\cdot X(t+\tau)\;dt = \mathbf{E}\{X(t)\cdot X(t+\tau)\} \tag{7.3}$$

The frequency is randomly varying around an average frequency $\bar{\nu}$, the instantaneous frequency $\nu(t)$ follows from the time derivative of the argument of the cosine term according to:

$$\nu(t) \;=\; \frac{1}{2\pi}\frac{d}{dt}\left(2\pi\bar{\nu}t \;+\; \phi(t)\right) \;=\; \bar{\nu} \;+\; \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{7.4}$$

As can be seen from equation (7.4), the time variable phase factor $\phi(t)$ fully accommodates the frequency bandwidth $\Delta\nu$ of the stochastic wave signal.

The above mathematical description suffices for a linearly polarized signal, however in

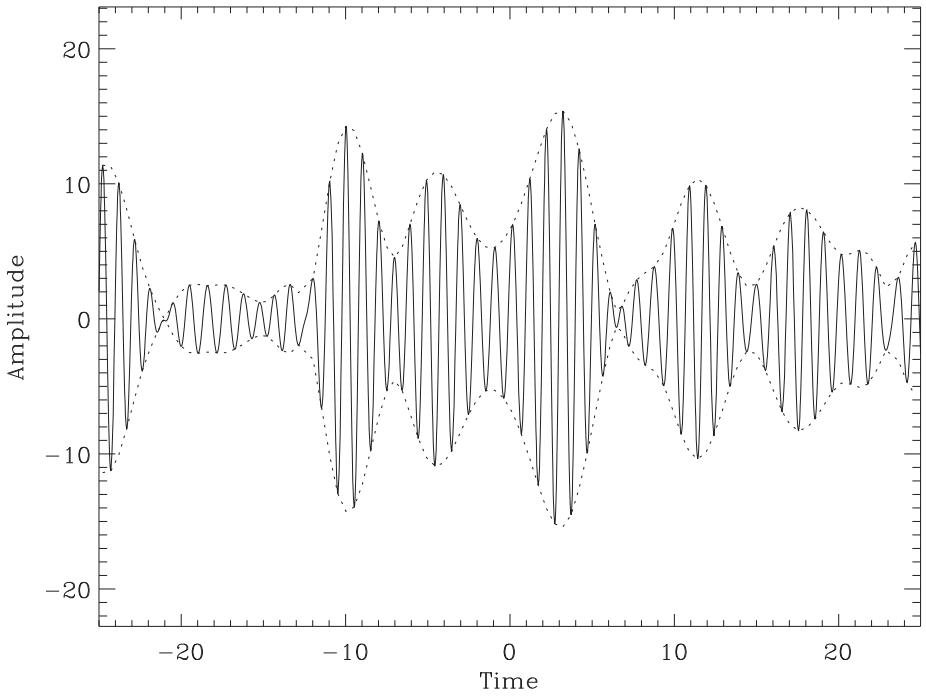


Figure 7.1: *A quasi-monochromatic EM-wave (one degree of polarization).*

case of a thermal radiator a particular *polarization direction is only very short-lived*, i.e. only during the *coherence time τ_c of an emitted wave packet*. This can be understood

by considering the emission processes involved.
Thermal emission consists of an extremely large number of radiative transitions, generated by randomly oriented atomic emitters. Each atom radiates a polarized wave train for roughly 10^{-8} or 10^{-9} seconds in the case of optical light depending on the natural line width $\Delta\nu$ of the transition. In the case of molecular vibrational or rotational transitions (radio and far infrared) the timescales are substantially longer. Considering a certain wave propagation direction $\vec{k}$, individual atomic (molecular) emissions along that direction will combine to generate a polarized wave, which however will not persist for more than the typical coherence time τ_c of a wave packet, i.e. in the optical $10^{-8} - 10^{-9}$ seconds. New wave trains are continually emitted and as a result the magnitude and the polarization direction of the electric vector $\vec{E}(t)$ changes in a completely random manner on a typical time scale equal to the coherence time τ_c. If these changes occur at a rate of 10^8 to 10^9 per second, any persistent polarization state is indiscernible. Thermal radiation is therefore designated as *natural or unpolarized light*, although the latter qualification is somewhat confusing since in actuality the light comprises a rapid succession of different polarization states.
The *rapid random fluctuations* in the electric vector $\vec{E}(t)$ of a thermal radiation field can be handled mathematically in a *scalar approach* by using a *complex analytic representation* of the quasi-monochromatic wave field.
Consider the time-varying electric field $E(t)$ of equation (7.1). Along with $E(t)$ one may consider a complex function:

$$\tilde{E}(t) = E(t) + iF_{Hi}(t) \tag{7.5}$$

in which:

$$F_{Hi}(t) \;=\; -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{E(t')dt'}{t'-t} \tag{7.6}$$

is the Hilbert transform of $E(t)$. This integral can be interpreted as a convolution of $E(t)$ with $(\pi t)^{-1}$:

$$F_{Hi}(t) \;=\; E(t) \,*\, \frac{1}{\pi t} \tag{7.7}$$

Applying the convolution theorem and considering the Fourier transform of $(\pi t)^{-1} \Leftrightarrow i\,\mathrm{sgn}(\nu)$, the Hilbert transform can be regarded as a *special filter* that leaves the amplitude of the spectral components unimpaired, but alters their phases by $\pi/2$, positively or negatively depending on the sign of ν. A consequence of this is, that Hilbert transforms of even functions are odd and those of odd functions even.
The complex function $\tilde{E}(t)$ of equation (7.5) is known as the *analytic signal*, the Hilbert transform is referred to as the *quadrature function* of $E(t)$. For example, the quadrature function of $\cos t$ is $\sin t$, the analytic signal is therefore $\exp(it)$.
Analytic functions are useful to describe quasi-monochromatic wave phenomena, where one deals with *modulated carrier* signals. The analytic signal contains *no negative frequency* components, it is obtainable from $E(t)$ by suppressing the negative frequencies and doubling the result.
For example $\cos 2\pi\nu_0 t$ contains frequency components at ν_0 and $-\nu_0$:

$$\cos 2\pi\nu_0 t = \frac{e^{2\pi i\nu_0 t} + e^{-2\pi i\nu_0 t}}{2}, \tag{7.8}$$

the analytic signal follows from suppression of $e^{-2\pi i\nu_0 t}$ and multiplication by 2.
If $E(t)$ is a Gaussian process, its Hilbert transform (linear) is also a Gaussian process, moreover the autocorrelation functions are equal and the values of $E(t)$ and its Hilbert transform are *uncorrelated* at the same instant t. The analytic signal comprises a harmonic oscillation at an average frequency $\bar{\nu}$ modulated by a slowly varying envelope:

$$\tilde{E}(t) = \tilde{E}_0(t) \cdot e^{i(2\pi\bar{\nu}t)} \tag{7.9}$$

The complex amplitude (envelope function) $\tilde{E}_0(t)$ can be expressed as:

$$\tilde{E}_0(t) = |\tilde{E}_0(t)| \cdot e^{i\phi(t)} \tag{7.10}$$

This envelope function is also referred to as the *phasor* of the analytic signal, $|\tilde{E}_0(t)|$ represents the instantaneous amplitude of $\tilde{E}(t)$ and $\phi(t)$ the time variable phase. The time rate of change of $\phi(t)$, i.e. $\left(\frac{1}{2\pi}\right)\frac{d\phi(t)}{dt}$, represents the instantaneous frequency shift $\Delta\nu(t)$ of the analytic signal relative to the average carrier frequency $\bar{\nu}$.

7.1.2 Non-linear mixing element: heterodyne detection

In many applications the signal power is often a very small fraction of the system noise. Hence, large values of the product $T_{obs}\cdot\Delta\nu$ are required to obtain the proper value of the limiting sensitivity for the detection of a weak source. As a consequence, the receivers employed must be very stable and well calibrated. Stable amplifiers are available at relatively low radio frequencies (i.e. up to 10 GHz), but large amplification factors and stability do become a problem at high frequencies. In order to overcome this problem, extensive use is made in radio and sub-millimeter applications of *heterodyne techniques*. In this technique, the incoming radio signal is mixed with a reference signal at a frequency relatively close to the average signal frequency $\bar{\nu}_s$. This reference signal at frequency ν_l is produced by a so-called local oscillator with high frequency stability, i.e. $\Delta\nu_l$ is almost a δ-function. In that case the reference signal has a very large coherence time and may therefore be regarded as coherent in the mixing process. Both signals are combined and fed to a non-linear detection element, the mixer element, which has the effect to produce signal power at both sum $\nu_s + \nu_l$ and difference $|\,\nu_s - \nu_l\,|$ frequencies (recall: $\nu_s = \frac{1}{2\pi}\frac{d}{dt}(2\pi\bar{\nu}_s t + \phi(t))$). This can be quantitatively described by considering the following processing sequence. The wave signal received at the focus of the telescope is first passed through a horn and a resonance cavity, which select the polarisation direction and the frequency bandwidth $\Delta\nu$. Subsequently the signal of the (tunable) local oscillator, at frequency ν_l, is coupled (superimposed) to the high frequency source signal ν_s through a so-called diplexer (or coupler) and funneled onto the non-linear mixer element. As discussed previously, a thermal source signal can be expressed as

$$\tilde{E}_s(t) = |\,\tilde{E}_{s,0}(t)\,| \cdot e^{i(2\pi\bar{\nu}_s t + \phi(t))} \tag{7.11}$$

Similarly, the signal of the local oscillator can be expressed as

$$\tilde{E}_l(t) = \tilde{E}_{l,0}(t) \cdot e^{i(2\pi\nu_l t)} \tag{7.12}$$

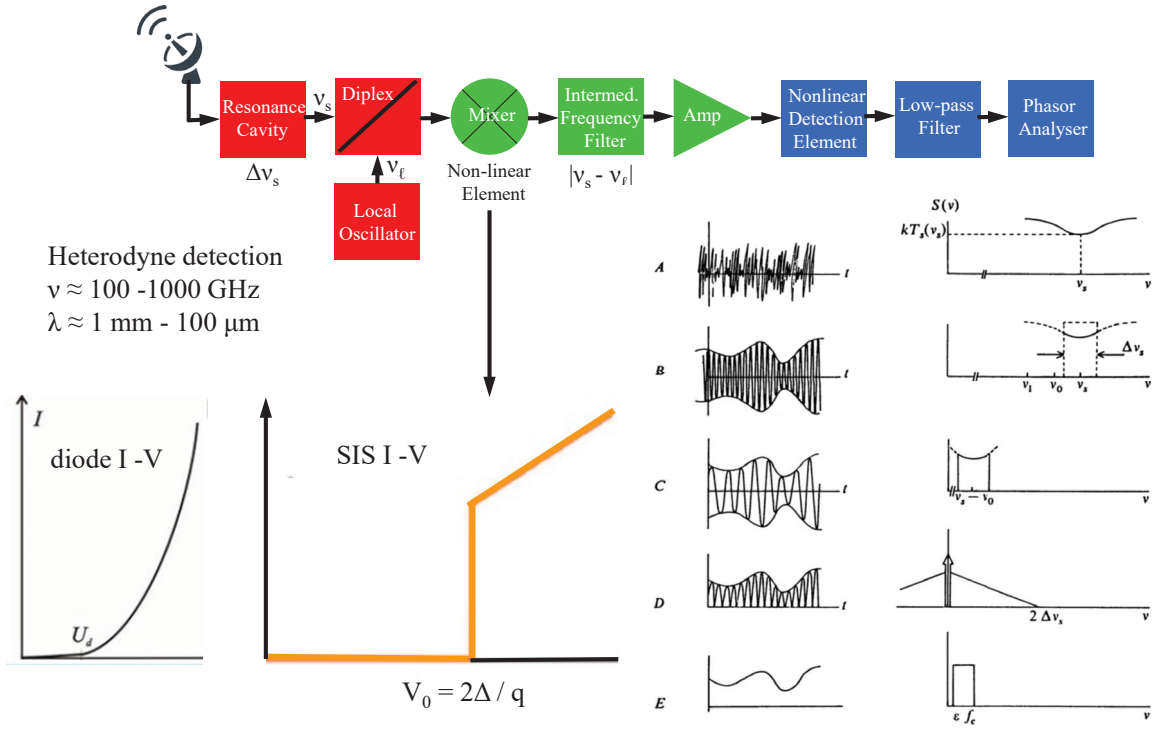


Figure 7.2: *Super-heterodyne processing chain.*

in which $\tilde{E}_{l,0}(t)$ represents the phasor of the local oscillator signal. Regarding the local oscillator as a coherent signal source, this phasor can be written as

$$\tilde{E}_{l,0}(t) = |\,\tilde{E}_{l,0}\,| \cdot e^{i\psi_l} \tag{7.13}$$

Hence:

$$\tilde{E}_l(t) = |\,\tilde{E}_{l,0}\,| \cdot e^{i(2\pi\nu_l t + \psi_l)} \tag{7.14}$$

Since the source signal concerns one component of polarization, the wave signal at the non-linear mixer element can be written as

$$E_M(t) = |\,\tilde{E}_{s,0}(t)\,| \cdot \cos(2\pi\bar{\nu}_s t + \phi(t)) + |\,\tilde{E}_{l,0}\,| \cdot \cos(2\pi\nu_l t + \psi_l) \tag{7.15}$$

Detection by the mixer element causes a *frequency conversion* through the occurrence of a product component arising from the non-linear response. Assuming a non-linear mixer response **proportional to** $E_M^2(t)$, a product component $2\ |\,\tilde{E}_{s,0}(t)\,| \cdot |\,\tilde{E}_{l,0}\,| \cdot \cos(2\pi\bar{\nu}_s t + \phi(t)) \cdot \cos(2\pi\nu_l t + \psi_l)$ is formed. Applying the trigonometric relation $\cos\alpha\ \cos\beta = \frac{1}{2}[\cos(\alpha+\beta)+\cos(\alpha-\beta)]$ two frequency components can be distinguished:

- one centered at $\bar{\nu}_s + \nu_l$: $|\,\tilde{E}_{s,0}(t)\,| \cdot |\,\tilde{E}_{l,0}\,| \cdot \cos(2\pi(\bar{\nu}_s + \nu_l)t + \phi(t) + \psi_l)$
- one centered at $|\bar{\nu}_s - \nu_l|$: $|\,\tilde{E}_{s,0}(t)\,| \cdot |\,\tilde{E}_{l,0}\,| \cdot \cos(2\pi(\bar{\nu}_s - \nu_l)t + \phi(t) + \psi_l)$

The latter expression is linearly proportional to the amplitude of the source signal $|\,\tilde{E}_{s,0}(t)\,|$ and possesses all the characteristics of the original signal, but now centers at a much lower difference frequency $|\,\bar{\nu}_s - \nu_l\,|$ (except for the phase shift ψ_l). In the

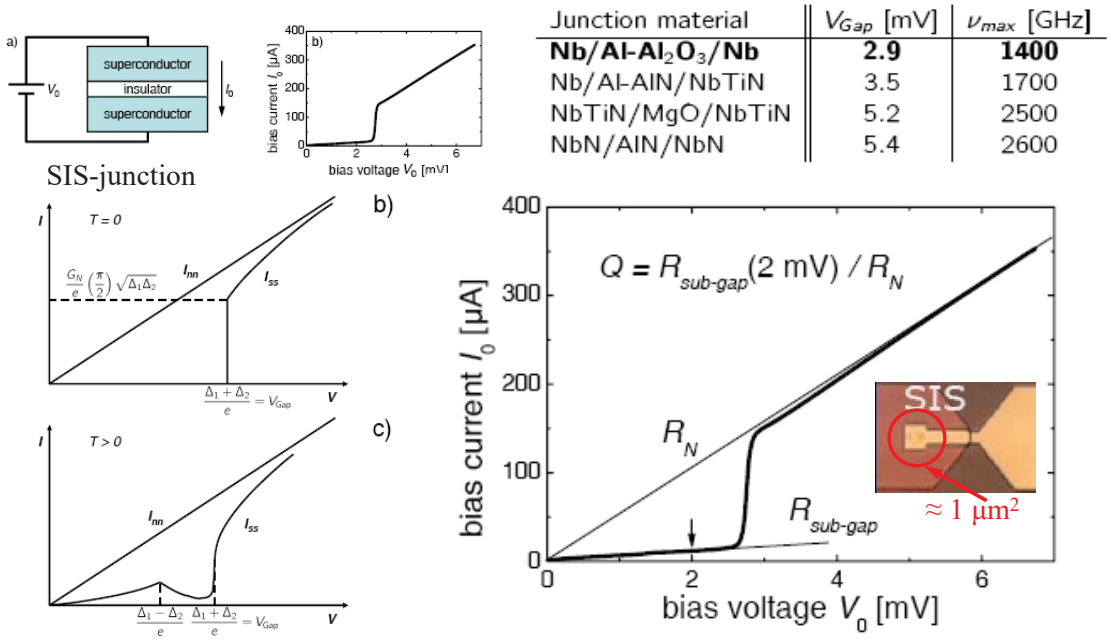


Figure 7.3: Characteristics of SIS junctions.

sub-millimeter application with $\bar{\nu}_s \approx 1$ THz, $|\bar{\nu}_s - \nu_l|$ is of the order 10 GHz. Remember that the bandwidth of the source signal $\Delta\nu_s$ is contained in the time dependent phase factor $\phi(t)$ and, consequently, the *mixer element should have adequate wide frequency response* to cover the complete bandwidth.
The converted signal at the difference or intermediate frequency (IF) can be selected by employing an intermediate frequency filter (IFF) and can subsequently be amplified by a stable low noise IF amplifier (A). Obviously the mixer element should have the lowest possible noise temperature, near-quantum limited operation near 1000 GHz has been achieved (see further).

By employing a tunable local oscillator the average frequency $\bar{\nu}_s$ of the source signal can be shifted while maintaining a fixed $\bar{\nu}_{IF}$ and associated filter circuitry, i.e.:

$$\bar{\nu}_s = \bar{\nu}_l\ (tunable)\ + \bar{\nu}_{IF}\ (fixed) \tag{7.16}$$

The IF signal is finally fed into a non-linear detection element (D) which, after low-pass filtering (LPF), generates the frequency spectrum associated with the signal phasor for one component of polarization $|\ \tilde{E}_{s,0}(t)\ |\ \ \cos\phi(t)$, i.e. the envelope function of the high frequency source signal $\tilde{E}_s(t)$. This frequency spectrum represents the spectral distribution of the radio signal centered at $\bar{\nu}_s$ over the bandwidth $\Delta\nu_s$. This signal can be fed into a frequency analyser (spectrometer) for spectral analysis. Figure (7.2) shows the signal processing chain and various stages in the detection process. Panel A displays the high frequency thermal signal incident on the telescope (antenna). Panel B shows the output of the resonance cavity into the diplexer, limited to the bandwidth $\Delta\nu_s$. Panel C shows the envelope function superimposed on the IF-carrier signal $\bar{\nu}_{IF} = |\bar{\nu}_s - \nu_l|$, after frequency conversion by the non-linear mixing element. In panel D this signal is fed through the non-linear detection element (normally a quadratic response) after which

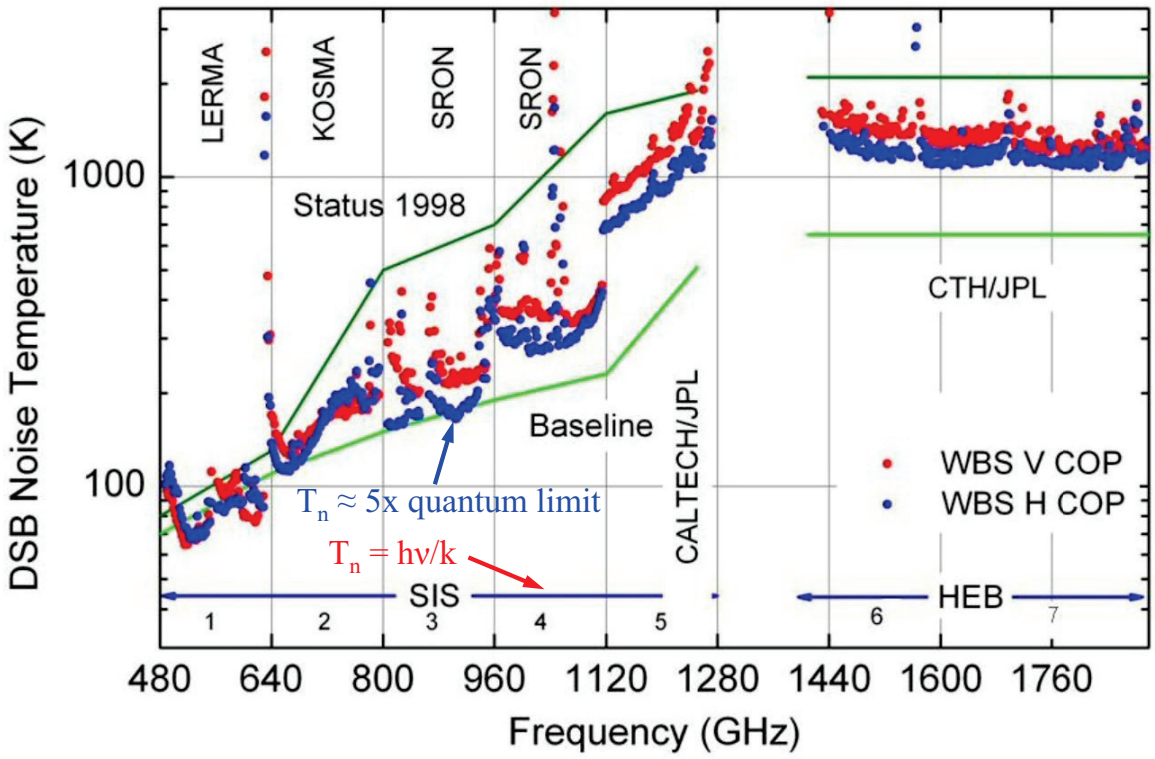


Figure 7.4: *In-flight noise temperatures achieved in several frequency channels of the HIFI-Wide Band Spectrometer on board the Herschel Space Observatory.*

the low frequency phasor signal can be filtered out (panel E). Heterodyne detection at sub-millimeter and infrared wavelengths is a powerful spectroscopic technique. In this case the mixer element consists of a *superconductive tunnel junction*, comprising two superconducting electrodes separated by a thin oxide barrier (see figure (7.3). Because of an energy gap 2Δ in the material of the superconducting electrodes, the current-voltage characteristics (the so-called I-V diagram) of this Superconductor Insulator Superconductor (SIS) junction is highly non-linear. SIS-mixers are the most sensitive heterodyne mixers in the 0.2–1.2 THz frequency range, where the upper frequency is determined by the energy gap 2Δ of practical superconducting materials (e.g. Niobium (Nb) and Niobium-Titanium (NbTi)). The frequency down-conversion process can be achieved with near quantum-limited noise performance. An example is shown in figure (7.4) that shows the noise temperatures which were obtained in a number of high frequency channels of the Wide Band Spectrograph (WBS) in the HIFI-instrument on board the Herschel Space Observatory. Up to 700 GHz a noise temperature of a few times (≈ 5) the quantum limit has been obtained, the mixer operates near the quantum limit, but input losses and IF amplifier noise decreases the overall performance. Above 1.2 THz hot-electron bolometer mixers (HEBM) are used which are not limited by the energy bandgap of the superconductor, since they do not operate on the individual photon energy but on the total absorbed power. The physical foundation of bolometer operation will be treated in a later section.

Note: The factor $|\tilde{E}_{l,0}|$ in the product component can be made large so that the signal

to noise ratio can be largely improved with respect to the internal noise sources in the receiver chain. This assumes of course that $|\tilde{E}_{l,0}|$ is ultra stable and, also, that the quantum noise of the local oscillator is still negligible.

7.1.3 Non-linear detection element: limiting sensitivity in the thermal limit

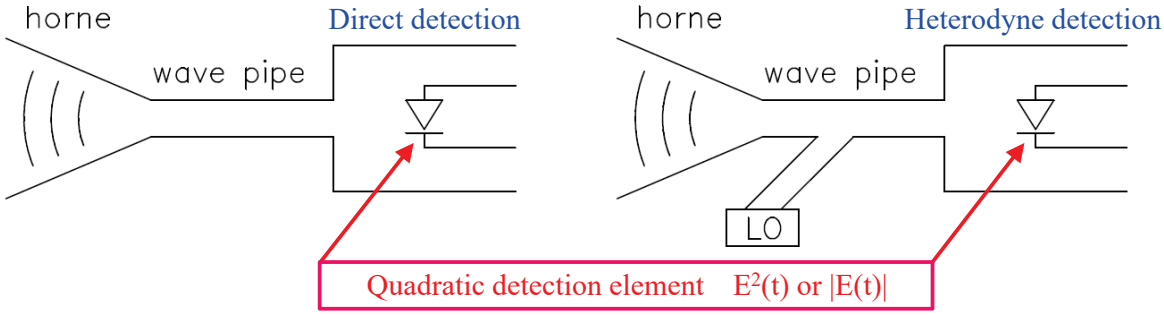


Figure 7.5: *Schematic view of a radio receiver with the receiving horn, which selects the frequency ν_s and the frequency bandwidth $\Delta\nu_s$, the wave pipe, and the 'quadratic' detection element (left). By mixing the radio signal with that of a local oscillator (LO), the carrier frequency can be shifted to much lower frequencies without any information loss (heterodyne detection), suitable for electronic processing (right).*

The average spectral noise power in the thermal limit ($h\nu \ll kT$) for one degree of polarization, equals kT. This is the situation which prevails in radio-, microwave- and sub-millimeter receivers and hence gives rise to a description of signals and noise with the aid of characteristic temperatures, e.g. source temperature T_s and noise temperature T_n. The one-sided power spectral density $S(\nu) \equiv \overline{P}(\nu) = kT$ Watt Hz^{-1} is constant as a function of frequency and, as a consequence, is termed white noise.
In practice, due to the finite frequency response of any receiver system, this will be frequency limited. Hence we can express the double-sided power spectral density for a thermal source for one degree of polarization as:

$$S_d(\nu) = \frac{1}{2}kT\Pi\left(\frac{\nu}{2\nu_c}\right) \quad \text{with} \quad \nu_c \ll \frac{kT}{h} \tag{7.17}$$

where ν_c constitutes the cut-off frequency of the receiver system under consideration, i.e. the total energy contained in the power spectrum remains finite, as it should be for any physical system.
Continuing now with the detection of radio signals, we wish to consider the signal-to-noise ratio. As discussed previously,the linearly polarized signal displayed in figure (7.1) can mathematically be expressed by the real function:

$$E(t) = E_0(t)\cos(2\pi\bar{\nu}t + \phi(t)) \tag{7.18}$$

The amplitude $w(t) \equiv E_0(t)$ of the quasi-monochromatic wave is a *wide-sense stationary Gaussian* random time function of zero mean. Moreover the stochastic process is assumed to be *mean- and correlation-ergodic*, i.e. for an arbitrary real stochastic variable $w(t)$ its *expectation value* at time t, $\mathbf{E}\{w(t)\}$, can be interchanged with its time average.
Detection of such a radio signal requires, like in the case of *mixing*, a non-linear operation like $\psi(t) \equiv |w(t)|$ or $\psi(t) = w^2(t)$, to extract the power present in the signal. We shall consider here the case of quadratic detection, i.e. $\psi(t) = w^2(t)$, since this is mathematically straightforward in contrast to absolute-value transformations.

If $w(t)$ can be described as a stationary random process with a normally distributed amplitude around zero mean, the probability density function is given by:

$$f(w) = \frac{1}{\sigma_w\sqrt{2\pi}} e^{-w^2/2\sigma_w^2} \qquad (\mu_w = 0) \tag{7.19}$$

and the measuring process is schematically indicated by

$$w(t) \rightarrow \text{transformation} \rightarrow \psi(t) \equiv w^2(t) \tag{7.20}$$

For the transformation $\psi(t) = w^2(t)$, with $\sigma_w^2 = R_w(0)$, we can write the probability density of ψ as

$$f(\psi) = \frac{1}{2[2\pi R_w(0)\psi]^{1/2}} \exp\left[\frac{-\psi}{2R_w(0)}\right] U(\psi) \tag{7.21}$$

with $U(\psi)$ the Heaviside step function. (In understanding expression (7.21), don't forget the transformation $dw \Rightarrow d\psi$!). Thus, the stochastic process $\psi(t)$ is apparently *not* normally distributed, and of course also $\mu_\psi \neq 0$.

To derive the power spectral density of $\psi(t)$ we need to find an expression for the autocorrelation function $R_\psi(\tau)$ of $\psi(t)$. One can show that for a normally distributed $w(t)$ the autocorrelation of $\psi(t) \equiv w^2(t)$ follows from:

$$\begin{aligned} R_\psi(\tau) &= \mathbf{E}\{\psi(t)\psi(t+\tau)\} = \mathbf{E}\left\{w^2(t)w^2(t+\tau)\right\} = \\ &= \mathbf{E}\left\{w^2(t)\right\}\mathbf{E}\left\{w^2(t+\tau)\right\} + 2\mathbf{E}^2\{w(t)w(t+\tau)\} \end{aligned} \tag{7.22}$$

The derivation of this relation applying the theory of stochastic processes makes use of the so-called moment-generating functions of $w(t)$. Hence:

$$R_\psi(\tau) = R_w^2(0) + 2R_w^2(\tau) = \mu_\psi^2 + C_\psi(\tau) \tag{7.23}$$

The average μ_ψ of $\psi(t)$ equals the variance of $w(t)$, the autocovariance of $\psi(t)$ equals twice the square of the autocovariance of $w(t)$, and the variance of $\psi(t)$ is $\sigma_\psi^2 = 2\sigma_w^4$. The *double-sided* power spectral density of $\psi(t)$ follows from the Wiener-Khinchin theorem:

$$S_{d_\psi}(\nu) = R_w^2(0)\delta(\nu) + 2S_{d_w}(\nu) * S_{d_w}(\nu) \tag{7.24}$$

It is important to realize that in the case of quadratic detection a number of frequency components is introduced in the case of an amplitude-modulated signal like $w(t)$. We

shall demonstrate this, as an example, for a deterministic signal, i.e. an amplitude-modulated high frequency carrier of the form:

$$x(t) = A(1 + m\cos\eta t)\cos\omega t \tag{7.25}$$

This represents a high frequency carrier (angular frequency ω) the amplitude (A) of which is modulated by a much lower frequency (angular frequency η). m is called the modulation index.
The signal $x(t)$ contains three discrete angular frequencies, $\omega - \eta$, ω, and $\omega + \eta$. This is easily seen by using $\cos\alpha\cos\beta = \frac{1}{2}[\cos(\alpha+\beta) + \cos(\alpha-\beta)]$.

The instantaneous intensity of this signal can be expressed as $I = x^2(t)$, and for the average intensity we get

$$\overline{I} = \overline{x^2(t)} = \frac{1}{2}A^2(1 + \frac{1}{2}m^2) \tag{7.26}$$

The equivalent of the term $\mathbf{E}\{a^2(t)\}\,\mathbf{E}\{a^2(t+\tau)\}$ in expression (7.22) amounts in this case to a DC-term representing the square of the average intensity:

$$\overline{I}^2 = \frac{1}{4}A^4\left(1 + \frac{1}{2}m^2\right)^2 \tag{7.27}$$

To arrive at the equivalent of the autocovariance term $2\mathbf{E}^2\{a(t)a(t+\tau)\}$ in equation (7.22), we first have to compute the autocovariance $C_x(\tau)$ of $x(t)$:

$$C_x(\tau) = \overline{x(t)x(t+\tau)} = \frac{1}{2}A^2[\cos\omega\tau + \frac{1}{4}m^2\cos(\omega+\eta)\tau + \frac{1}{4}m^2\cos(\omega-\eta)\tau] \tag{7.28}$$

Subsequently:

$$\left[\overline{x(t)x(t+\tau)}\right]^2 = \frac{1}{4}A^4(1 + \frac{1}{2}m^2\cos\eta\tau)^2\cos^2\omega\tau \tag{7.29}$$

Using the relation $\cos^2\alpha = \frac{1}{2}(\cos 2\alpha + 1)$ this can be disentangled in various components:

$$\begin{aligned}\left[\overline{x(t)x(t+\tau)}\right]^2 &= \frac{A^4}{8}(1+\cos 2\omega\tau) + \frac{m^2A^4}{8}[\cos\eta\tau + \frac{1}{2}\cos(2\omega-\eta)\tau + \\ &+ \frac{1}{2}\cos(2\omega+\eta)\tau] + \frac{m^4A^4}{64}[1 + \cos 2\eta\tau + \frac{1}{2}\cos(2\omega-2\eta)\tau + \\ &+ \cos 2\omega\tau + \frac{1}{2}\cos(2\omega+2\eta)\tau]\end{aligned} \tag{7.30}$$

This expression shows that the original angular frequencies in the power spectral density of $x(t)$, ω and $\omega \pm \eta$ have been transformed in the squaring process to angular frequency components at DC, η, 2η, 2ω, $2\omega \pm \eta$ and $2\omega \pm 2\eta$. Displayed on a double-sided angular frequency diagram this shows twice the original angular frequency bandwidth centered at DC and $\pm 2\omega$ (see figure 7.7). In the case of the stochastic signal $w(t)$ we shall see the same characteristics in the power spectral density of the autocovariance $C_\psi(\tau)$.

Consider now the front-end of a receiver behind a radio antenna (telescope), see (Figure (7.1). This front-end generally contains a resonance cavity tuned at a central frequency $\overline{\nu} = \nu_s$ with a bandwidth $\Delta\nu_s$. If the noise entering the receiver can be

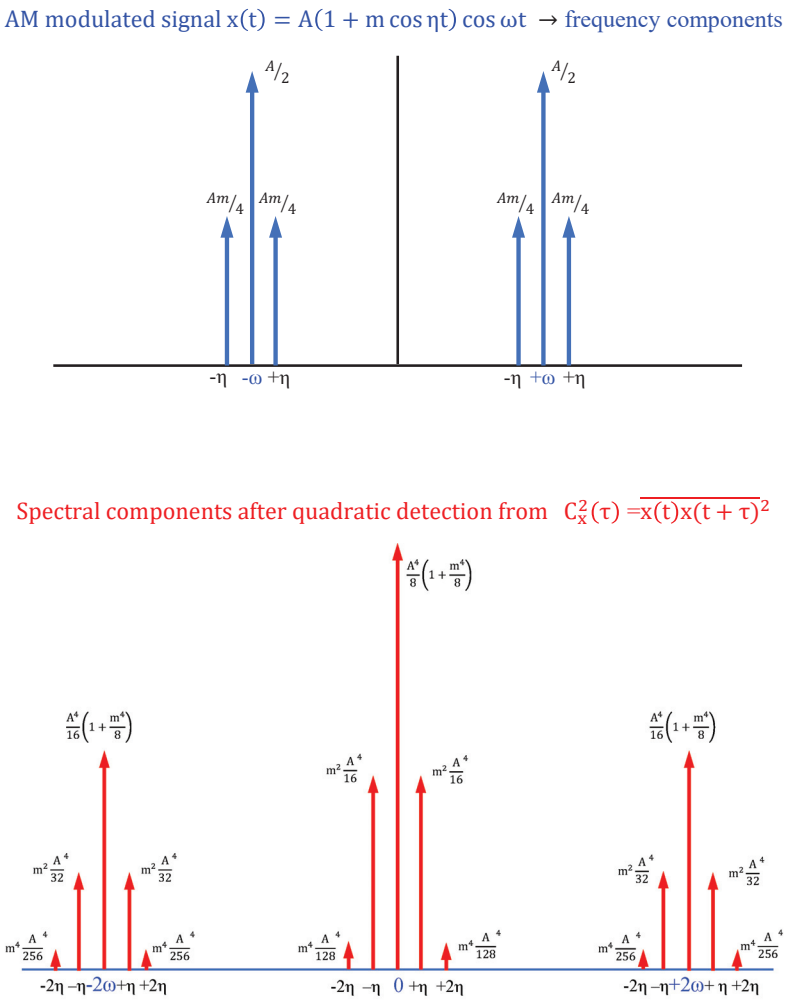


Figure 7.6: *Frequency components of an amplitude modulated (angular frequency η) carrier (angular frequency ω) with modulation index m (upper panel). The spectral components that emerge after quadratic detection present in the autocovariance function are shown in the lower panel. Evidently the sidebands appear at twice the carrier angular frequency ω, moreover the signal bandwidth 4η amounts to twice the original bandwidth 2η.*

characterized by a noise temperature T_n the double-sided power spectral density of $w(t)$ for one degree of polarization is given by:

$$S_{d_w}(\nu) = \frac{1}{2} k T_n \left[\Pi\left(\frac{\nu - \nu_s}{\Delta\nu_s} \right) + \Pi\left(\frac{\nu + \nu_s}{\Delta\nu_s} \right) \right] \tag{7.31}$$

This signal is then fed to a non-linear detection element, like a Schottky-diode or an induction coil, which introduces the transformation $\psi(t) = w^2(t)$. Consequently, we

have:

$$R_w^2(0) = (\sigma_w^2)^2 = \left(\frac{1}{2} kT_n \int_{-\infty}^{+\infty} \left[\Pi\left(\frac{\nu - \nu_s}{\Delta\nu_s} \right) + \Pi\left(\frac{\nu + \nu_s}{\Delta\nu_s} \right) \right] d\nu \right)^2 = (kT_n \Delta\nu_s)^2 \quad (7.32)$$

and

$$\begin{aligned} 2\left[S_{d_w}(\nu) * S_{d_w}(\nu)\right] &= \frac{1}{2}(kT_n)^2 \left[\Pi\left(\frac{\nu - \nu_s}{\Delta\nu_s} \right) + \Pi\left(\frac{\nu + \nu_s}{\Delta\nu_s} \right) \right] * \left[\Pi\left(\frac{\nu - \nu_s}{\Delta\nu_s} \right) + \Pi\left(\frac{\nu + \nu_s}{\Delta\nu_s} \right) \right] \\ &= (kT_n)^2 \Delta\nu_s \left[\Lambda\left(\frac{\nu}{\Delta\nu_s} \right) + \frac{1}{2}\Lambda\left(\frac{\nu - 2\nu_s}{\Delta\nu_s} \right) + \frac{1}{2}\Lambda\left(\frac{\nu + 2\nu_s}{\Delta\nu_s} \right) \right] \end{aligned} \quad (7.33)$$

$S_{d_\psi}(\nu)$ consists therefore of a component $(kT_n\Delta\nu_s)^2\delta(\nu)$, a *time independent average value at zero frequency* (stationary Gaussian random amplitudes) and three 'triangle-functions' centered at zero frequency and at frequencies $-2\nu_s$ and $+2\nu_s$ with a base width of $2\Delta\nu_s$, as illustrated in figure 7.7. Note the correspondence in frequency shift and bandwidth with the example involving the deterministic amplitude modulated signal above! In practice one always has $\nu_s \gg \Delta\nu_s$, in the centimeter range for example

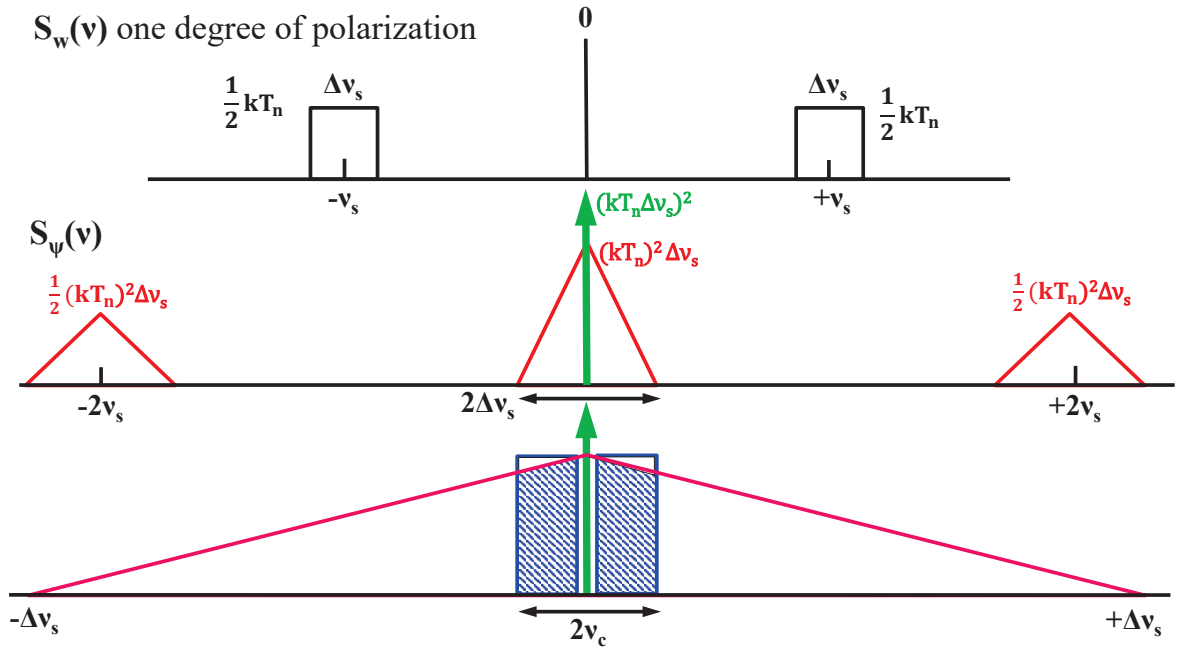


Figure 7.7: *Upper panel: Double-sided power spectral density of thermal radiation for one degree of polarization ($\bar{P} = kT$ Watt Hz^{-1}), over a channel bandwidth $\Delta\nu_s$ centered at frequency ν_s, incident on a non-linear detection element. Middle panel: double-sided power spectral density at the output of the detection element. Lower panel: low frequency filtering (cut-off ν_c) providing signal averaging over a time interval $\Delta T_{av} = 1/2\nu_c$. This time averaging process obviously does not influence the DC-component, this is schematically indicated in the figure by showing an exclusion of the green arrow.*

one typically has $\nu_s \simeq 10^{10}$ Hz and $\Delta\nu_s \simeq 10^7$ Hz. The detection of a *potential radio*

source signal will then have to be assessed in the context of the *autocovariance of the noise signal*, i.e. we need an expression for $C_\psi(\tau)$. This follows from the Fourier transform of the term $2\,[S_{d_w}(\nu) * S_{d_w}(\nu)]$ in equation (7.33). Applying the shift theorem and the $\Lambda \Leftrightarrow \mathrm{sinc}^2$ transform from Fourier analysis, we get:

$$\begin{aligned} C_\psi(\tau) &= (kT_n\Delta\nu_s)^2\left[\frac{e^{-2\pi i(2\nu_s\tau)}+e^{2\pi i(2\nu_s\tau)}}{2}\right]\mathrm{sinc}^2\tau\Delta\nu_s + (kT_n\Delta\nu_s)^2\mathrm{sinc}^2\tau\Delta\nu_s \\ &= (kT_n\Delta\nu_s)^2(1+\cos 2\pi(2\nu_s\tau))\mathrm{sinc}^2\tau\Delta\nu_s \end{aligned} \tag{7.34}$$

The $\cos 2\pi(2\nu_s\tau)$ term refers to the high frequency carrier which will be filtered off in any low frequency averaging process. This averaging process can be taken over an arbitrary time intervalΔT_{av}. This is equivalent to filtering in the frequency domain with a filter $\Pi(\nu/2\nu_c)$ with $\nu_c = 1/(2\Delta T_{av})$ commensurate with the Nyquist sampling theorem.
The averaged value of the autocovariance is then obtained in the τ domain by convolution of $C_\psi(\tau)$ with the Fourier transform $\Pi(\nu/2\nu_c) \Leftrightarrow 2\nu_c\mathrm{sinc}2\nu_c\tau$.
Assuming $\nu_c \ll \Delta\nu_s \ll \nu_s$, the $\cos 2\pi(2\nu_s\tau)$ term in expression (7.34) averages to zero. Consequently we have:

$$\begin{aligned} [C_\psi(\tau)]_{\Delta T} &= (kT_n\Delta\nu_s)^2\mathrm{sinc}^2\tau\Delta\nu_s * 2\nu_c\mathrm{sinc}2\nu_c\tau \\ &= (kT_n\Delta\nu_s)^2\cdot 2\nu_c\int_{-\infty}^{+\infty}\mathrm{sinc}^2\tau'\Delta\nu_s\mathrm{sinc}2\nu_c(\tau-\tau')d\tau' \end{aligned} \tag{7.35}$$

and with a change of variables $u' \equiv \tau'\Delta\nu_s$ this becomes:

$$[C_\psi(u)]_{\Delta T} = (kT_n)^2\Delta\nu_s\cdot 2\nu_c\int_{-\infty}^{+\infty}\mathrm{sinc}^2u'\mathrm{sinc}\frac{2\nu_c}{\Delta\nu_s}(u-u')du' \tag{7.36}$$

Since $\nu_c/\Delta\nu_s \ll 1$, $\mathrm{sinc}(2\nu_c/\Delta\nu_s)u'$ varies very slowly compared to sinc^2u'. We may therefore regard sinc^2u' as a δ-function in comparison to $\mathrm{sinc}(2\nu_c/\Delta\nu_s)u'$. Moreover we also have the proper normalization, since $\int_{-\infty}^{\infty}\mathrm{sinc}^2u'du' = \int_{-\infty}^{+\infty}\delta(u')du' = 1$. Applying this approximation we get:

$$\begin{aligned} [C_\psi(u)]_{\Delta T} &= (kT_n)^2\Delta\nu_s\cdot 2\nu_c\int_{-\infty}^{+\infty}\delta(u')\mathrm{sinc}\frac{2\nu_c}{\Delta\nu_s}(u-u')du' \\ &= (kT_n)^2\Delta\nu_s\cdot 2\nu_c\mathrm{sinc}\frac{2\nu_c}{\Delta\nu_s}u \end{aligned} \tag{7.37}$$

Substituting $\tau = u/\Delta\nu_s$ we arrive at the final expression for the ΔT-averaged value of the noise autocovariance:

$$[C_\psi(\tau)]_{\Delta T} = (kT_n)^2\Delta\nu_s\cdot 2\nu_c\mathrm{sinc}2\nu_c\tau \tag{7.38}$$

The noise variance $[C_\psi(0)]_{\Delta T} = (kT_n)^2 \Delta\nu_s \cdot (2\nu_c)$ should be compared to the strength of a radio source signal characterized by a source temperature T_s. The average value of this source signal after quadratic detection follows from:

$$(\mu_\psi)_s = R_{w_s}(0) = \sigma^2_{w_s} = kT_s\Delta\nu_s \tag{7.39}$$

and the signal-to-noise ratio thus becomes:

$$\mathrm{S/N} = \frac{(\mu_\psi)_s}{[C_\psi(0)]^{1/2}_{\Delta T}} = \frac{T_s}{T_n}\left(\frac{\Delta\nu_s}{2\nu_c}\right)^{1/2} \tag{7.40}$$

i.e. the signal to noise is proportional to the square root of the receiver bandwidth $\Delta\nu_s$ and inversely proportional to the double-sided bandwidth of the integrating (averaging) low-pass filter ν_c. Introducing $\Delta T_{av} = 1/(2\nu_c)$ (Nyquist sampling) we get:

$$\mathrm{S/N} = \frac{T_s}{T_n}\left(\Delta\nu_s \Delta T_{av}\right)^{1/2} \tag{7.41}$$

i.e. the S/N ratio improves with the square root of the product of the radio-channel bandwidth and the integration time ΔT_{av}. In practice the noise temperature of a radio-wave receiving system is designated as the system or operational temperature that includes the contributions to the noise of the sky, the antenna and the receiver system. The S/N =1 sensitivity for detecting a radio source against the system thermal noise temperature is sometimes referred to as the *radiometer equation*:

$$\mathrm{S/N} = \frac{T_{source}}{T_{system}}\left(\Delta\nu_s \Delta T_{av}\right)^{1/2} \quad \text{radiometer equation!} \tag{7.42}$$

The minimum detectable *source power for one degree of polarization* with a signal to noise ratio of one is given by:

$$(P_s)_{min} = kT_n\left(\Delta\nu_s \Delta T_{av}\right)^{-1/2} \tag{7.43}$$

7.2 Photoconductive element

7.2.1 Operation principle and Responsivity

A photoconductor exhibits a change in conductance (resistance) when radiant energy (photons) is incident upon it. The radiant energy increases the conductance (σ_0) by producing *excess charge carriers* in the photoconductor, which is fabricated from a semiconductor material. The charge carriers comprise electron-hole pairs in the case of an intrinsic semiconductor. In extrinsic semiconductors the excess charge carriers comprise either electrons (n-type) or holes (p-type). The spectral responsivity of a particular photoconductor is determined by its energy gap: only photons that have energies greater than the gap energy will be absorbed and cause an excess current flow. The photoconductor is operated in a mode that involves the application of an external electric field (voltage bias) that gives rise to a bias current. This bias current becomes

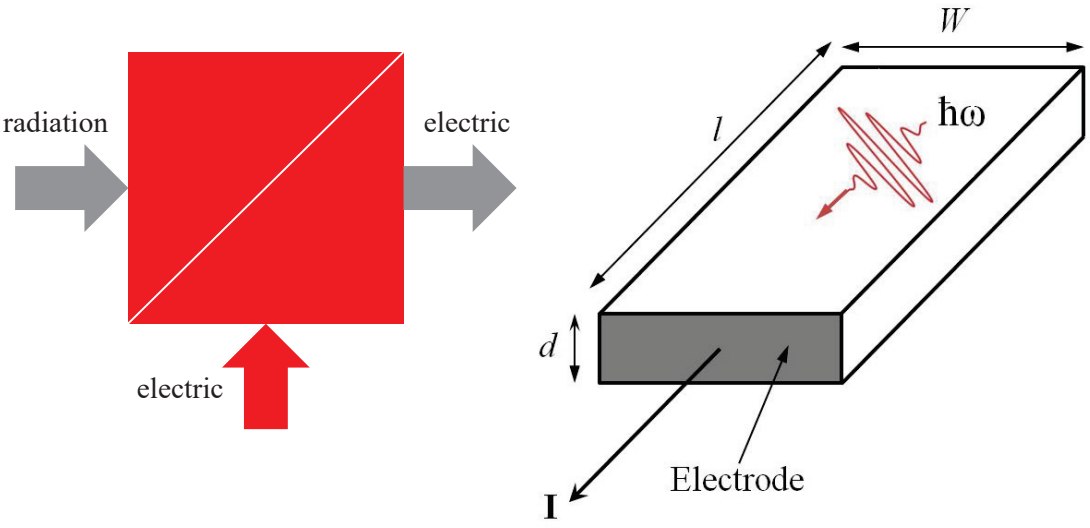


Figure 7.8: *(Left): Photoconductors belong to the category of modulation transducers. (Right): Sketch of the geometry of a photoconductive element.*

modulated by the excess charge carriers that are produced by photon-excitation. Hence a photoconductor belongs to the category of modulation transducers (see figure (7.8). To put the above discussion about the operation principle on a proper physical basis, the response to a radiative signal will be derived in this section. A material with conductivity σ_0 produces a current density $\vec{j}$ given by:

$$\vec{j} = \sigma_0 \vec{E} \tag{7.44}$$

in which $\vec{E}$ represents the electric field strength generated by the bias voltage V_B across the photoconductor ($\mid \vec{E} \mid$ in Volt$\cdot$m^{-1} and σ_0 in Ohm$^{-1}\cdot$m^{-1}). Microscopically the current density $\vec{j}$ can also be expressed as

$$\vec{j} = Nq\vec{v} \tag{7.45}$$

in which N represents the volume density of the free charge carriers, q the elementary charge and $\vec{v}$ the drift velocity of these charges in the applied electric field. This drift velocity $\vec{v}$ can also be written as $\vec{v} = \mu_c \vec{E}$, with μ the so-called mobility of the charge carrier. For an intrinsic semiconductor a distinction needs to be made between electron conduction and hole conduction: the mobilities μ_n for electrons and μ_p for holes are quite different ($\mu_n \approx 3\mu_p$):

$$\vec{j} = -nq\vec{v_n} + pq\vec{v_p} \tag{7.46}$$

with n/p the electron/hole densities and $\vec{v_n}/\vec{v_p}$ the electron/hole drift velocities (opposite directions). Note that q is the elementary charge and has a positive sign. Combining equations (7.44) and (7.46), an expression for the conductivity σ_0 follows:

$$\sigma_0 = q(n\mu_n + p\mu_p) \tag{7.47}$$

which reduces to $\sigma_0 = qn\mu_n$ and $\sigma_0 = qp\mu_p$ in the case of a heavily doped n-type, respectively p-type extrinsic semiconductor.
Consider now a n-type semiconductor, which is irradiated by a beam of radiant energy with a spectral photon irradiance (= monochromatic photon flux density) $F(\lambda)$. In the equilibrium situation the number of excess conduction electrons follows from the fact that the generation rate should match the recombination rate:

$$\frac{d\Delta n}{dt} = g - \frac{\Delta n}{\tau_l} = 0 \tag{7.48}$$

with Δn the equilibrium number of excess electrons per unit volume (= excess carrier concentration), τ_l the life time of these electrons against recombination and g the generation rate:

$$g = \frac{\eta(\lambda)F(\lambda)}{d} \tag{7.49}$$

with $\eta(\lambda)$ the photon detection efficiency and d the thickness of the photoconductor material. The increase in conductivity $\Delta\sigma = \sigma - \sigma_0$ follows from:

$$\Delta\sigma = q\mu_n\Delta n = \frac{q\mu_n\eta(\lambda)F(\lambda)\tau_l}{d} = \frac{q\mu_n\eta(\lambda)\tau_l}{Ad}\,\frac{\lambda}{hc}\,\Phi(\lambda) \tag{7.50}$$

in which $\Phi(\lambda)$ the monochromatic radiation flux in Watt, A the illuminated area of the photoconductor and λ the wavelength under consideration. Given a fixed bias voltage V_B across the photoconductors, the relative change in conductivity $\Delta\sigma/\sigma$ can be related to a relative change in current $\Delta I/I_0$ and resistance $\Delta R/R_0$ as:

$$\frac{\Delta\sigma_0}{\sigma_0} = -\frac{\Delta R}{R_0} = \frac{\Delta I}{I_0} = \frac{I_{pc}}{I_0} \tag{7.51}$$

in which I_0, R_0 represent the photoconductor DC-current/resistance in the absence of radiation and $\Delta I = I_{pc}$ the photon-generated current (*photo-current*).
Introducing the detector width W and length l ($A = lW$, see figure (7.8), I_0 can be expressed as

$$I_0 = Wd\,|\vec{j}\,| \tag{7.52}$$

and hence:

$$I_{pc} = I_0\,\frac{\Delta\sigma_0}{\sigma_0} = \frac{\eta(\lambda)q\lambda}{hc}\cdot\frac{\tau_\ell\mu_n V_B}{l^2}\cdot\Phi(\lambda) \tag{7.53}$$

The term $\tau_\ell\mu_n V_B/l^2$ is commonly referred to as the photoconductive gain G. Introducing the transition time τ_{tr} of the free charge carriers accross the photoconductor length l, i.e. $\tau_{tr} = \frac{l^2}{\mu_n V_B}$, G gives the ratio between the carrier life time against recombination in the photoconductor and its transition time, i.e. $G = \tau_\ell/\tau_{tr}$.
The *current responsivity* $R^I_{pc}(\lambda,\nu)$, assuming a modulation of the photon flux density $\Phi(\lambda)$ with frequency ν follows from:

$$R^I_{pc}(\lambda,\nu) = \frac{I_{pc}}{\Phi(\lambda)} = G\eta(\lambda)q\,\frac{\lambda}{hc} \tag{7.54}$$

in Ampere/Watt.
In practice the photocurrent I_{pc} is measured over a load resistance R_L in series with

the photoconductor resistance R_0, translating I_{pc} in an equivalent voltage V_0, see figure (7.9). In this case the *spectral voltage responsivity* is given by:

$$R^V_{pc}(\lambda,\nu) = \frac{R_L R_0}{R_L + R_0}\, G\eta(\lambda) q\, \frac{\lambda}{hc} \tag{7.55}$$

in Volt/Watt. ◁

The above expressions are only valid for low-frequency operation of the photoconductor, where dielectric relaxation or recombination life-time effects are not of concern. In the next section, the frequency response of photoconductors will be addressed.
In order to raise the responsivity of a photoconductor, one should

- enlarge the quantum efficiency $\eta(\lambda)$ by minimizing reflection effects at the entrance plane through application of anti-reflection coatings and by creating a larger cross-section for the internal photo-electric effect.
- increase the carrier life time, which raises the photoconductive gain G.
- enlarge the carrier mobility μ_c (recall that n-type charge carriers have substantially higher mobility than p-type charge carriers).
- increase the operational bias voltage V_B, which lowers the value of the transition time τ_{tr} in the expression for the photoconductive gain G.

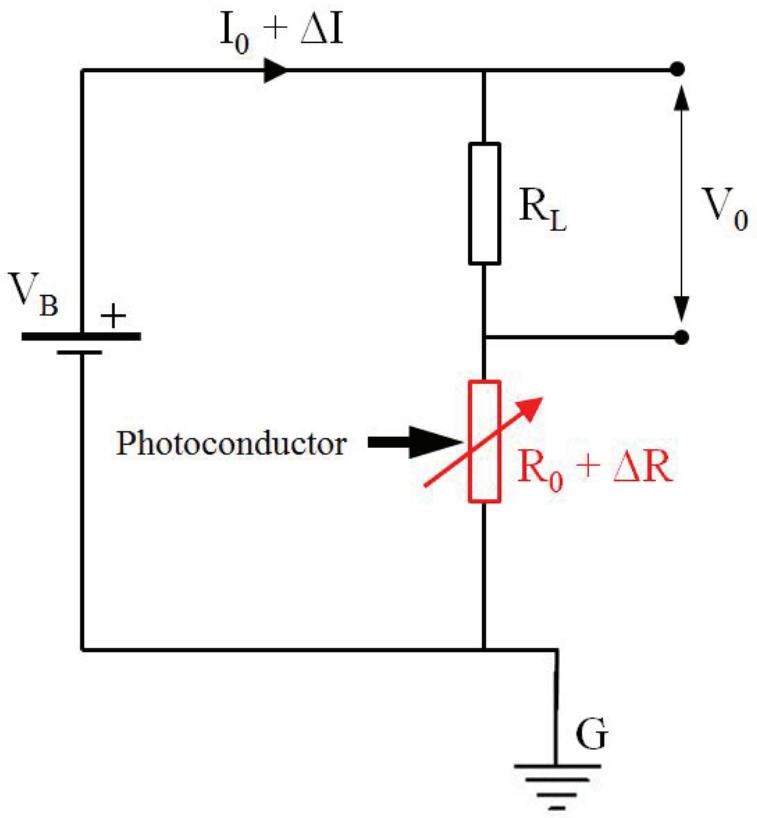


Figure 7.9: *Photoconductor: bias circuit.*

7.2.2 Temporal frequency response

Consider a step-function in charge carrier generation due to a switch-on of the photoconductor exposure to a radiation source. The response of the photoconductor is now obtained from solving the time dependent continuity equation:

$$\frac{d\Delta n(t)}{dt} = g - \frac{\Delta n(t)}{\tau_\ell} \tag{7.56}$$

with boundary condition $\Delta n = 0$ at $t = 0$. The solution of this differential equation is straightforward:

$$\Delta n(t) = g\tau_\ell(1 - e^{-\frac{t}{\tau_\ell}}) = \Delta n_{eq}(1 - e^{-\frac{t}{\tau_\ell}}), \tag{7.57}$$

with $\Delta n_{eq} = g\tau_\ell$ the *equilibrium* excess free carrier density [m^{-3}] The time dependent

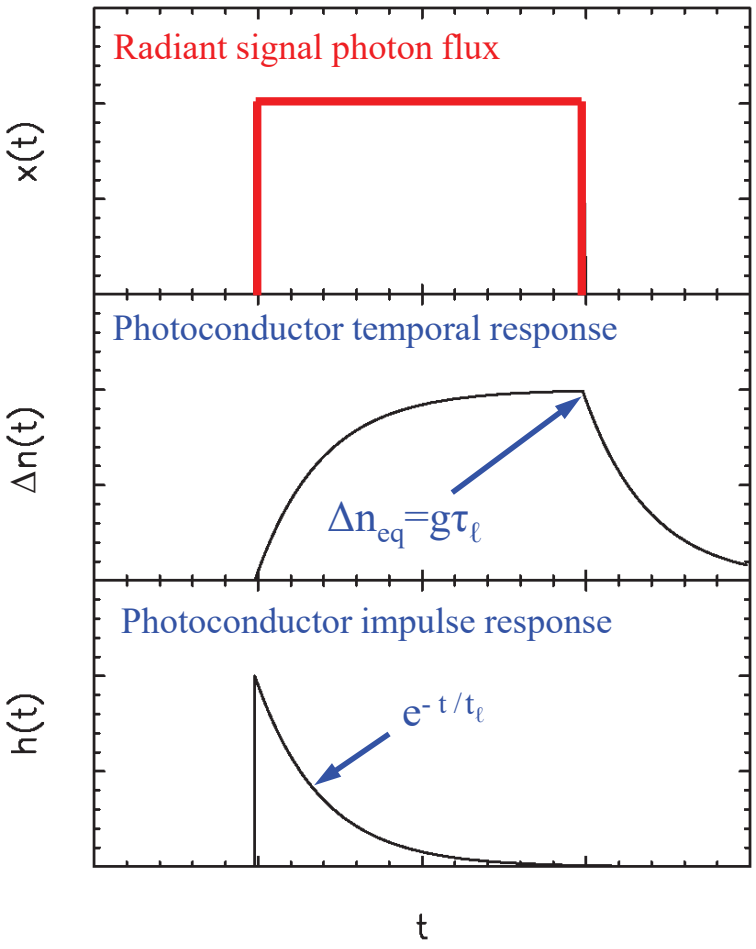


Figure 7.10: *For a block function $x(t)$ of illuminating photons (top), the number of excess charges per unit volume [m^{-3}] $\Delta n(t)$ in a photoconductor increases on a finite time scale towards the equilibrium value, and drops exponentially once the illumination ceases (middle). The impulse response function $h_{pc}(t)$ is an exponential (below).*

excess carrier concentration (density) can also be expressed as the convolution of a

charge generation step-function, $\Delta n_{eq}U(t)$ with the photoconductor impulse response function $h_{pc}(t)$:

$$\Delta n(t) = \Delta n_{eq}U(t) * h_{pc}(t) \tag{7.58}$$

Fourier (Laplace) transform for $t \geq 0$ to the temporal frequency domain yields:

$$\Delta\bar{n}(2\pi j\nu) = \frac{\Delta n_{eq}}{2\pi j\nu}\,\bar{H}_{pc}(2\pi j\nu) \tag{7.59}$$

Substituting expression (7.57) in $\Delta n(t) \Leftrightarrow \Delta\bar{n}(2\pi j\nu)$ yields the photoconductor response function for a dynamical input signal:

$$\frac{\Delta n_{eq}}{2\pi j\nu}\,\frac{1}{1+2\pi j\nu\tau_\ell} = \frac{\Delta n_{eq}}{2\pi j\nu}\,\bar{H}_{pc}(2\pi j\nu) \qquad \Rightarrow \qquad \bar{H}_{pc}(2\pi j\nu) = \frac{1}{1+2\pi j\nu\tau_\ell} \tag{7.60}$$

and hence the frequency dependent amplitude transfer:

$$|\,\bar{H}_{pc}(2\pi j\nu)\,| = \frac{1}{\sqrt{1+4\pi^2\nu^2\tau_\ell^2}} \tag{7.61}$$

This expression shows that the photoconductor acts as a frequency filter that takes the form of a first order system (see figure 7.10). At a frequency $\nu = 1/(2\pi\tau_\ell)$, the responsivity reduces with 3 dB ($= 1/\sqrt{2}$). For $2\pi\nu\tau_\ell \gg 1$, the responsivity rolls off with 6 dB per octave; i.e. a reduction of a factor 2 by each doubling of the frequency. The cut-off frequency $\nu_c = 1/(2\pi\tau_\ell)$ determines the characteristic bandwidth of a photoconductor with a charge carrier life time τ_l. This breaking point in the frequency domain is typically around 1 MHz, which of course varies for various detector types. Typical response times of photoconductive radiation sensors are, therefore, intrinsically limited to microsecond time scales.

7.2.3 GR-noise in the *signal-photon-limit*

Ideally, the noise related to the photon irradiance on the active area of the sensor should determine the noise, this constitutes the best possible situation. To arrive at that situation, the GR-noise (photon noise) should be the dominant noise component, i.e. $\overline{\Delta I^2_{GR_{ph}}} \gg \overline{\Delta I^2_{th}} + \overline{\Delta I^2_{1/f}}$, where $\overline{\Delta I^2_{th}}$ represents the thermal noise component and $\overline{\Delta I^2_{1/f}}$ the $1/f$ noise contribution.
In the chapter on noise sources we derived a general expression for generation - recombination (GR) noise, i.e. expressions in (5.83). We shall now apply this to the specific case of a photoconductor.
The average number of charge carriers generated by a radiation beam with an average monochromatic photon flux density $F(\lambda)$ (average spectral photon irradiance) at wavelength λ equals $G\eta(\lambda)F(\lambda)A_{pc}$, in which A_{pc} represents the active area of the photoconductor, $\eta(\lambda)$ the quantum efficiency for photo-absorption and G the photoconductive gain. The average photocurrent $\mathbf{E}\{I_{ph}(t)\}$ can therefore be expressed as:

$$\mathbf{E}\{I_{ph}(t)\} = \overline{I_{ph}(t)} = qG\eta(\lambda)F(\lambda)A_{pc} \tag{7.62}$$

with q the elementary charge. The frequency response can be expressed as (see equation (7.60)):

$$\bar{H}_{pc}(2\pi j\nu) = \frac{1}{1 + 2\pi j\nu\tau_\ell} \tag{7.63}$$

where τ_ℓ is the charge carrier life time.
Proceeding here with the result of the derivation outlined in the chapter on noise sources, see equation (5.83), the GR-current noise can be expressed as:

$$\left(\sqrt{\overline{\Delta I^2}}\right)_{GR_{ph}} = \sqrt{4qG\overline{I_{ph}(t)}\Delta\nu_\ell} = 2qG\sqrt{\eta(\lambda)F(\lambda)A_{pc}\Delta\nu_\ell} \tag{7.64}$$

with $\left(\sqrt{\overline{\Delta I^2}}\right)_{GR_{ph}}$ the rms-noise current, $\overline{I_{ph}(t)}$ the average total photo-current, $F(\lambda)$ the average radiant signal photon flux and $\Delta\nu_\ell = \int_0^{+\infty} d\nu/(1 + [2\pi\nu\tau_\ell]^2)$ the one-sided noise equivalent bandwidth within the frequency range $0 < \nu < \infty$. Substituting $\Delta\nu_\ell$ by performing the integration over frequency we get:

$$\left(\sqrt{\overline{\Delta I^2}}\right)_{GR_{ph}} = qG\left(\frac{\eta(\lambda)F(\lambda)A_{pc}}{\tau_\ell}\right)^{\frac{1}{2}} \tag{7.65}$$

Finally, for the signal to noise ratio *in the signal photon limit* we obtain

$$\text{SNR} = \left(\frac{\overline{I_{ph}(t)}}{(\sqrt{\overline{\Delta I^2}})_{GR_{ph}}}\right) = (\eta(\lambda)F(\lambda)A_{pc}\tau_\ell)^{1/2} \tag{7.66}$$

This last equation tells us that a high value for the frequency cutoff $\nu_\ell = 1/(2\pi\tau_\ell)$ leads to a lower signal to noise for the photo-current; the reason for this is that the intrinsic system noise is less filtered.

7.2.4 GR-noise in the *background-photon-limit*: radiation BLIP normalized detectivity D^*

The rms-noise in the photocurrent, derived in the previous paragraph, dominates over thermal noise if the photoconductor is sufficiently cooled.
From expression (7.65) for the rms-noise in the photocurrent $\left(\sqrt{\overline{\Delta I^2}}\right)_{GR_{ph}}$ we can also assess the spectral Noise Equivalent Power ($NEP(\lambda, \Delta\nu_\ell)$) if we assume that this photocurrent arises from an *average monochromatic background-photon flux density* $B(\lambda)$ instead of the average signal-photon flux density $F(\lambda)$ considered above. This $NEP(\lambda, \Delta\nu_\ell)$ represents the so-called *radiation Background Limited Performance (BLIP)*. Let us use the relations:

$$NEP(\lambda, \Delta\nu_\ell) = \frac{\left(\sqrt{\overline{\Delta I^2}}\right)_{GR_{ph}}}{R_{I_{pc}}(\lambda, \Delta\nu_\ell)} = \frac{hc}{\lambda}\left(\frac{B(\lambda)A_{pc}}{\eta(\lambda)\tau_\ell}\right)^{\frac{1}{2}} \tag{7.67}$$

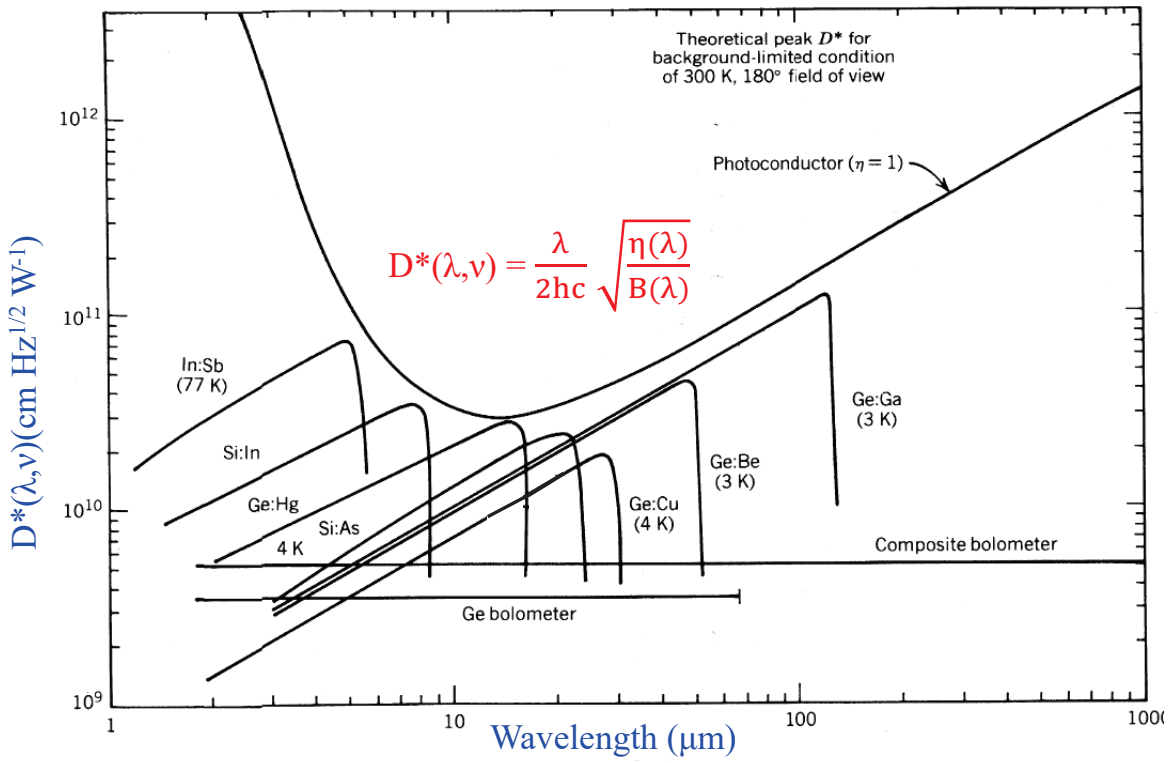


Figure 7.11: *Typical normalized detectivities for several photo-conductors. Credit Dereniak & Crow (1984) .*

and the BLIP normalized detectivity $D^*(\lambda, \Delta\nu_\ell)$:

$$D^*(\lambda, \Delta\nu_\ell) = \frac{\sqrt{A_{pc}\Delta\nu_\ell}}{NEP(\lambda, \Delta\nu\ell)} = \frac{\lambda}{2hc}\left(\frac{\eta(\lambda)}{B(\lambda)}\right)^{\frac{1}{2}} \quad \text{by substituting } \Delta\nu_\ell = 1/(4\tau_\ell) \tag{7.68}$$

Figure 7.11 shows the theoretical *peak detectivity* for $\eta(\lambda) = 1$ as a function of wavelength, assuming radiation Background Limited Performance(BLIP) arising from an omnidirectional black body radiation field at 300 K integrated over the upper hemisphere of 180°. Also included in figure (7.11) are a number of common extrinsic photo-conductors and their corresponding operating temperatures. Detectivities of two types of thermal detectors (bolometers) are displayed for comparison. The thermal detectors lack the cut-off feature in wavelength due to the different detection principle, they show a considerably lower value of $D^*(\lambda, \Delta\nu_\ell)$ but cover a wider spectral band than the photo-conductive devices.

7.2.5 'Dark' noise contributions

So far we have only considered the generation of GR-noise in photoconductive devices as a consequence of irradiation by an external photon source, both as a signal source or as an omnidirectional background source. However, if the photo-conductor is not

exposed to external radiation, there still remain intrinsic noise contributions owing to the presence of a resistive bias circuit, the thermal environment of the device and imperfections in the homogeneity of the conducting material and the electrical contacts. These noise components are often labelled as 'dark noise' contributions since they do not relate to photoconductor irradiance. The imperfections in materials and contacts cause the 1/f noise component that we briefly treated in an earlier chapter and since photoconductors require a bias current there will always be 1/f noise present. The mean square current noise is obtained from integrating the current power spectral density given in equation (5.94) over the equivalent noise bandwidth $\Delta\nu_{1/f}$:

$$\overline{\Delta I^2_{1/f}} = \int\limits_{\Delta\nu_{1/f}} S_I(\nu)d\nu \;=\; a\, I_b^2 \int\limits_{\Delta\nu_{1/f}} \frac{d\nu}{\nu} \tag{7.69}$$

with a a proportionality constant and I_b the bias current through the photoconductor. For a photoconductor it is an observational fact that at low frequencies the dominant noise exhibits a 1/f dependence incorporated in the expression for its current power spectral density. As the frequency increases this component drops below the GR-noise or the thermal noise in case of a low photon flux, see figure (7.12). The 1/f noise component does not constitute a fundamental limit to sensitivity, careful surface preparation and contact technology can potentially reduce the integral contribution of this noise to negligible levels. In addition to the GR-noise generated by an external radiation source, we also have a GR-noise contribution to the dark noise owing to thermal agitation of charge carriers:

$$\left(\sqrt{\overline{\Delta I^2}}\right)_{GR_{th}} = \sqrt{4q^2G^2g_{th}\Delta\nu_\ell} = qG\sqrt{\frac{g_{th}}{\tau_\ell}} \quad \text{with } \Delta\nu_\ell = 1/(4\tau_\ell) \tag{7.70}$$

in which g_{th} represents the thermal generation rate. If the device is sufficiently cooled, the thermal generation will decrease so that it can be neglected compared to the other noise contributions.
Thermal noise occurs in all resistive materials. If we assume that the photo-conductive device has a resistance R_{pc} and is part of a a bias circuit as displayed in figure (7.9), the mean square thermal noise current can be expressed as (see the chapter on noise sources):

$$\overline{\Delta I^2_{th}} = 4k\left(\frac{T_{pc}}{R_{pc}} + \frac{T_L}{R_L}\right)\Delta\nu_c = \frac{k}{\tau_n}\left(\frac{T_{pc}}{R_{pc}} + \frac{T_L}{R_L}\right), \text{ substituting } \Delta\nu_c = 1/(4\tau_n) \tag{7.71}$$

where T_{pc} and T_L are the temperatures of the photo-conductive sensor and the load resistor (R_L) respectively, τ_n represents the high frequency cut-off parameter of the electronic processing filter. For simplicity we shall assume that this electronic filter comprises a single integration with time constant τ_n and transfer function $\bar{H}(2\pi j\nu) = 1/(1+2\pi j\nu\tau_n)$. If the load resistor is mounted on the heat sink of the photoconductor, we have $T_{pc} = T_L = T$, which is advantageous for the thermal noise of the load resistor, since photoconductors are often cryogenically cooled. We can then substitute R_{pc} and R_L by a single parallel equivalent resistor $R_{par} = R_{pc}R_L/(R_{pc}+R_L)$.

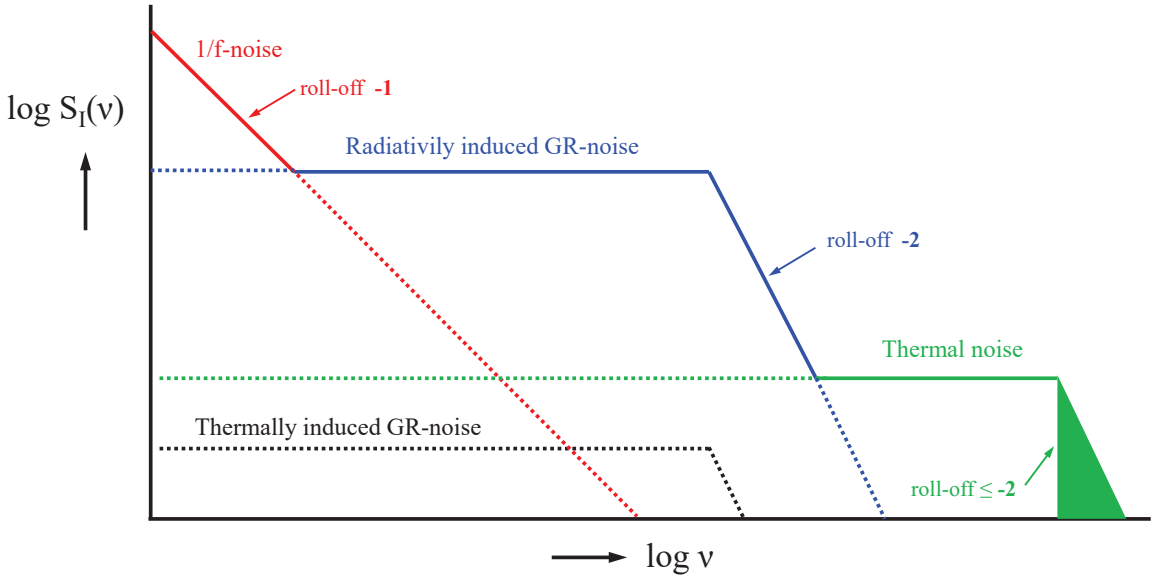


Figure 7.12: *Bode diagram of the current power spectral densities $S_I(\nu)$ for noise components encountered in photo-conductive devices as a function of temporal frequency. At low frequency the 1/f component dominates with a roll-off slope of -1 ($\propto \nu^{-1}$), between 100 Hz and 1 MHz, see figure (5.8) the GR-noise due to external irradiance should dominate if the temperature is kept low enough to suppress the thermally generated GR-noise (black dotted) and the thermal noise. Above the roll-off point due to the life time τ_ℓ of the radiation generated charge carriers (slope -2), the thermal noise will gradually take over at the higher frequencies up to a high frequency cut-off determined by the chosen filter pass band in the processing electronics (roll-off slope $\leq$ -2).*

All noise components add in quadrature (summation of noise variances), hence for the total mean square noise current, including both BLIP noise and 'dark' noise, we can now write:

$$\overline{\Delta I_{tot}^2} = \left(\frac{a\, I_b^2 \Delta\nu_{1/f}}{\nu} + \frac{kT}{R_{par}\tau_n} + \frac{q^2G^2}{\tau_\ell}[\eta(\lambda)B(\lambda)A_{pc} + g_{th}] \right) \tag{7.72}$$

The current power spectral densities of these various noise components are displayed in figure (7.12). As already mentioned, the total contribution of the $1/f$ noise can be made negligible by proper component manufacturing technologies and cooling will significantly suppress the thermally generated GR-noise. Therefore, in practice, the thermal resistor noise in the bias circuit will be the main factor in determining to which irradiance level background-photon GR-noise will dominate the photoconductor's performance, i.e we have the condition:

$$\frac{T}{R_{par}} \ll \frac{q^2G^2}{k}\left(\frac{\tau_n}{\tau_\ell}\right)[\eta(\lambda)B(\lambda)A_{pc}] \tag{7.73}$$

In practice the ratio $\tau_n/\tau_\ell \approx 1$ so as to arrive at an optimum match between the electronic filter and the photoconductor response that is governed by the charge carrier

lifetime.
In order to remain background-photon-noise(BLIP)-limited, the ratio of temperature to resistance must go down as $B(\lambda)$ decreases, so this sets certain limits to the load resistor and the operational temperature of the photo-conductive element. For high performance extrinsic IR sensors, the load resistor is much smaller than the sensor resistance, so $R_{par} \approx R_L$ allowing for external tuning. However for intrinsic photo-conductors like HgCdTe $R_{par} \approx R_{pc}$, in that case resistance tuning in the bias circuit is no option.
Equation (7.68) shows that the BLIP spectral D^* gets better if the level of the background irradiance goes down. If it is reduced to zero no further improvement in the value of D^* can be achieved since the condition given in equation (7.73) is no longer fulfilled. This is the case where the thermal noise becomes the dominant noise contribution. Hence, we can write the following expression for the spectral noise equivalent power $NEP(\lambda, \Delta\nu)$:

$$NEP(\lambda, \Delta\nu) = \frac{hc}{\lambda q G \eta(\lambda)} \left(\frac{kT}{R_{par}\tau_n} \right)^{\frac{1}{2}} \tag{7.74}$$

From this we can now derive the D^* for the thermal-noise-limited case of a photo-conductive element:

$$D^*(\lambda, \Delta\nu_c) = \frac{\sqrt{A_s \Delta\nu_c}}{NEP(\lambda, \Delta\nu_c)} = \frac{\lambda}{2hc}[qG\eta(\lambda)] \left(\frac{A_s R_{par}}{kT} \right)^{\frac{1}{2}}, \quad \text{with } \Delta\nu_c = 1/(4\tau_n) \tag{7.75}$$

From this expression for the normalized detectivity in case of thermal-noise-limited performance, it can be seen that the maximum achievable D^* is governed by the resistance-area ($A_s R_{par}$) product, i.e. high performance photoconductors require a high RA-product. Moreover the performance is also linearly related to the photoconductive gain, so sensors with an intrinsically low RA-product, like HgCdTe, can still attain a relatively high D^* value owing to a high photoconductive gain factor. For sensors with an $R_s \gg R_L$ operating in the thermal noise limit, the theoretically achievable maximum D^* is determined by the value of the load resistor R_L and is described by the same equation as the one for a photovoltaic sensor in which the equivalent R_{par} is in the junction resistance (viz. the next paragraph on photovoltaic sensor elements). As an example: for the low-background case, in order for a photoconductive sensor with a $G = 0.5$ to have equivalent performance as the photovoltaic sensor, the load resistance R_L should be four times higher than the junction resistance of the photovoltaic sensor.

7.3 Photovoltaic element: photo-diode and photo-transistor

7.3.1 Operation principle and Responsivity

These sensors belong to the category of modulating transducers.
The photodiode comprises a pn-junction that is strongly reverse-biased, so only a re-

verse (saturation) current j_s is flowing, see equation(4.28). Figure (7.13) shows the electron potential and current densities in the reverse biased photodiode pn-junction. Far away from the space charge region that resides in the depletion layer, we have equilibrium concentrations of minority charge carriers in the p-type and n-type semiconductor material denoted by n_{pe} (p-material) and p_{ne} (n-material). As discussed before when treating the diffusion capacitance of a pn-junction, in the vicinity of the depletion layer these equilibrium concentrations of minority charge carriers drop exponentially to zero at the edge of the depletion zone when a large reverse bias is applied, see figure (4.7). This exponential drop has also been marked in figure (7.13), extending to the edge of the depletion layer at $|x| = x_0$. Thus we have for $x \geq x_0$ and $x \leq -x_0$:

$$n_p(x) = n_{pe}\left(1 - e^{(x_0 - x)/L_n}\right), x \geq x_0 \text{ and } p_n(x) = p_{ne}\left(1 - e^{(x_0 + x)/L_p}\right), x \leq -x_0 \quad (7.76)$$

The minority diffusion current densities arising from these concentration gradients at the boundary with the depletion zone follow from:

$$j_{dn} = -(-q)D_n \left|\frac{\partial n_p(x)}{\partial x}\right|_{x=x_0} = \left(\frac{qD_n}{L_n}\right) n_{pe} \quad (7.77)$$

$$j_{dp} = -(+q)D_p \left|\frac{\partial p_n(x)}{\partial x}\right|_{x=-x_0} = \left(\frac{qD_p}{L_p}\right) p_{ne} \quad (7.78)$$

where D_n and D_p are the electron/hole diffusion coefficients and L_n and L_p represent the electron/hole diffusion lengths respectively.

If the photodiode is irradiated with light of photon energy $h\nu \geq E_{gap}$, electron-hole pairs will be generated in the space charge region and beyond. Within the depletion layer electrons and holes will be immediately separated owing to the presence of the electric field, outside the depletion layer but within roughly one diffusion length (L_n and L_p) diffusive transport of n and p *minority* charge carriers will take place into the depletion layer, where they subsequently will drift in the E-field towards the n- and p-material respectively. The increase of the number of charge carriers is *independent* of the applied reverse-bias voltage level, new equilibrium concentrations n'_{pe} and p'_{ne} get established by a constant level of photon irradiance according to:

$$n'_{pe} = n_{pe} + n_{ph} \quad \text{and} \quad p'_{ne} = p_{ne} + p_{ph} \quad (\text{independent of } V_{ex}) \quad (7.79)$$

This results is a change of the reverse bias current with an amount $\Delta I = I_{ph}$. Figure (7.14) shows the photodiode as current source along with two current-voltage characteristics: one in the absence of light (I_{dark}) and one when illuminated by a radiation source (I_{light}), the magnitude of the photocurrent governed by the irradiance level. The light illumination effect causes the I-V characteristic of the diode to shift downward by an amount $I_{ph} = I_{light} - I_{dark}$. In the earlier treatment of the pn-junction we derived the following relation for I_{dark}, see equation(4.28):

$$I_{dark} = I_s\left(e^{qV_{ex}/kT} - 1\right) \text{ with } I_s \text{ the reverse-bias saturation current} \quad (7.80)$$

The saturation current is generated by the minority charge carriers that are also the source of the photocurrent I_{ph}, so for I_{light} we can write:

$$I_{light} = I_s\left(e^{qV_{ex}/kT} - 1\right) - I_{ph} = I_s\left(e^{qV_{ex}/kT}\right) - (I_s + I_{ph}) \text{ with } I_{ph} = q\eta F(\lambda)A_s \quad (7.81)$$

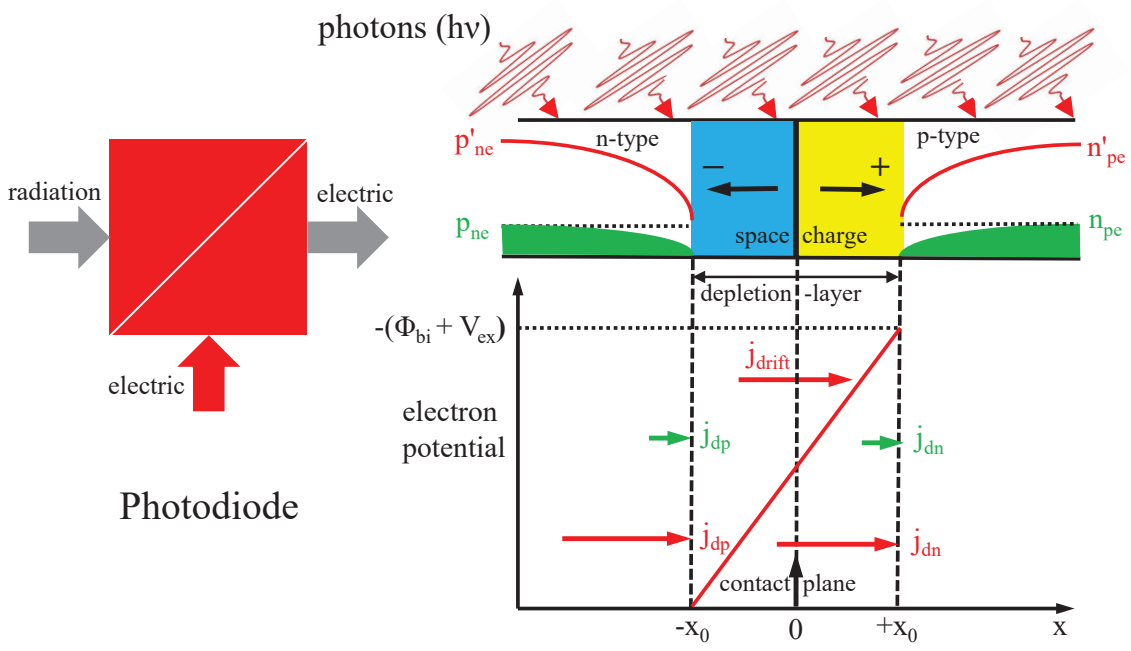


Figure 7.13: *Operational principle of a photodiode. The pn-junction is strongly reverse-biased giving rise to a large transition in electron potential that aids the passage of minority charge carriers across the barrier. The concentration gradients of the minority charge carriers in the absence of light are shown in green starting from equilibrium concentrations p_{ne} in the n-type material and p_{ne} in the p-type material. This gives rise to the minority diffusion current densities j_{dp} and j_{dn} (green arrows). When the device gets illuminated, these minority concentrations are raised in proportion to the incoming photon flux density (red curves), generating a drift current density j_{drift} when the interactions occur in the depletion layer (no recombination), and enhanced diffusion current densities j_{dp} and j_{dn} within a diffusion length of the depletion zone (red arrows).*

with η the detection efficiency (including reflection losses), $F(\lambda)$ the incident spectral photon flux density and A_s the irradiated active sensor area. Although the photodiode was introduced above as a current source, it can also be used as a power generator: the short circuit current delivered by the device is given by I_{ph}, the open circuit voltage V_{oc}, see the I-V characteristic in figure (7.14), is given by:

$$V_{oc} = \frac{kT}{q} ln \left(\frac{I_{ph} + I_s}{I_s} \right) \tag{7.82}$$

Power generation by photodiode devices as solar cells will be treated further on in this chapter.

Figure (7.14) also shows a schematic of a planar photodiode. This device has been configured in such a way that the p-type material can be maximally illuminated, whereas the n-type material is hardly reached due to photoelectric absorption of the incoming photon beam. This means that the properties are mainly determined by L_n, i.e. the minority charge carriers (electrons) in the p-type layer. Moreover, if the majority of the

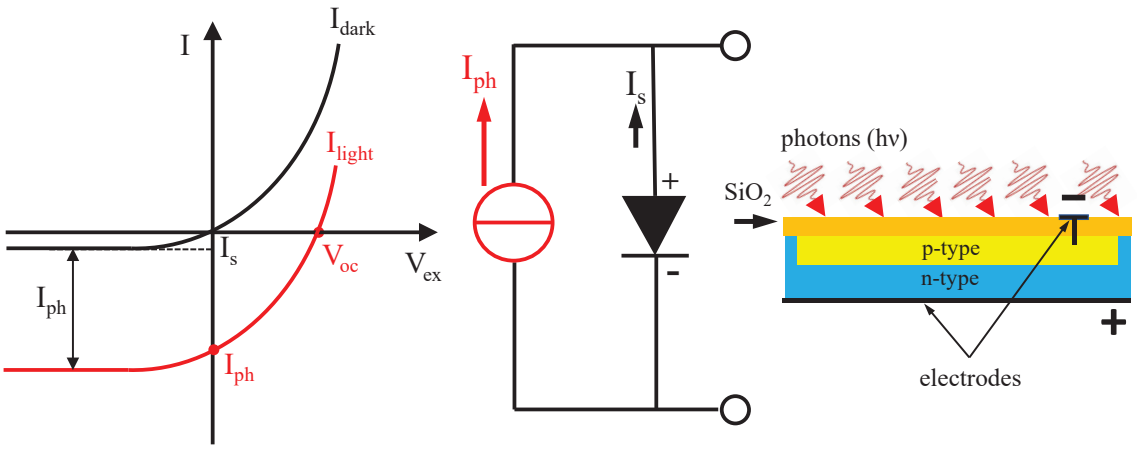


Figure 7.14: *The photodiode as a light driven current source. The two I-V curves refer to no illumination I_{dark} (black) and to light exposure I_{light} (red). The short circuit current I_{ph} and open circuit voltage V_{oc} points are indicated. A schematic layout of a planar photodiode is displayed in the right most picture.*

photon-induced electron-hole pairs can be produced in the depletion layer itself or in its close vicinity, recombination of the charge carriers can be neglected. Figure (7.15) shows the principle of a phototransistor, which is equipped with only two ohmic contacts, the base (B) is not electrically connected but is irradiated by light. The electron-potential diagram that is also displayed in figure (7.15) illustrates the electrical set-up for an npn-transistor. An external voltage (V_{ex}) is applied in such a way to the ohmic contacts that the collector-base (CB) junction is reverse-biased. The planar configuration shown maximizes the base area that can be irradiated. In the base layer the absorbed photons create electron-hole pairs, the minority charge carriers (electrons in this case) move immediately to the collector which is at a positive electrical potential (= negative electron-potential as shown in the figure). The holes that are produced cannot follow, since they are blocked by the potential barrier between B and C. As a consequence the base becomes charged-up to a more positive electrical potential which causes the emitter-base (EB) junction to move into forward bias. The electrons in E (majority charge carriers) flow into B and recombine there with the excess hole concentration. However, due to the much higher impurity concentration in the emitter material (more heavily doped), only a small fraction of the injected electrons recombines, the great majority reaches the collector aided by the electric field in the BC junction. This causes a substantial current amplification (β), just as in the case of a normal transistor in a common emitter circuit configuration:

$$I_C = I_{ph}(1 + \beta) \tag{7.83}$$

So, contrary to the situation with the photodiode, the phototransistor introduces a current amplification factor in the photovoltaic detection process. In fact it can be regarded as a photo-sensor with built-in amplification. However if we impose high demands on the noise characteristics or on the high frequency transit (pulse) response,

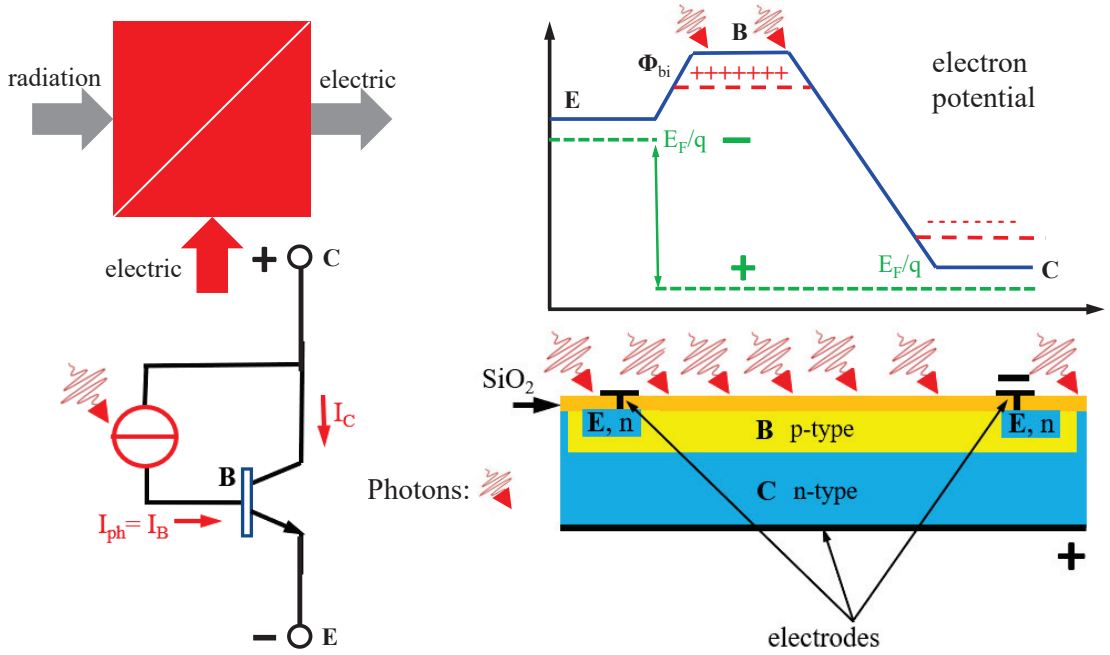


Figure 7.15: *Operation principle of a phototransistor. The voltage applied to the ohmic contacts connected to the emitter and collector (green plus and minus) is chosen in such a way that the base-collector junction is on reverse bias: blue lines in the electron potential diagram (upper right, npn-transistor). The transistor base is configured in a planar fashion (lower right) to maximize the light collecting power. The light incident on the base produces electron-hole pairs, the electrons escape to the collector aided by the electric field in the transition region. The holes built up an enhanced positive electric potential (red dashes) that moves the emitter-base junction further into forward bias leading to electron injection from the emitter into the base where a small fraction recombines and the bulk is transmitted to the collector. This leads to amplification of the photocurrent I_{ph} generated in the base layer (lower left).*

it is preferable to split the device in a photodiode and a separate transistor-amplifier (e.g. low-noise Field Effect Transistor amplifier) that can be independently optimized for optimum performance.
Photovoltaic devices are commonly operated as current generators. The spectral current responsivity is given by:

$$R^I_{pd}(\lambda,\nu) = \frac{I_{ph}}{\phi(\lambda)} = \frac{q\lambda\eta(\lambda)}{hc} = 0.808\lambda\eta(\lambda) \ \text{[Ampere Watt}^{-1}\text{]} \quad \text{with } \lambda \text{ in } \mu\text{m} \qquad (7.84)$$

As in the case of the photoconductor the responsivity is proportional to the wavelength up to a cut-off wavelength determined by the gap energy of the semiconductor in use. The proportionality to the wavelength follows from the convenience to measure the irradiance power in Watt, whereas the radiant energy conversion is based on the quantized interaction with the photons of the radiation beam. If we were to introduce the *photon*

spectral current responsivity $R^I_{pd_{ph}}$:

$$R^I_{pd_{ph}}(\lambda,\nu) = \frac{I_{ph}}{\Pi(\lambda)} = q\eta(\lambda) \tag{7.85}$$

which no longer exhibits the proportionality to the wavelength λ, $\Pi(\lambda)$ represents the total incident photon flux. The quantum efficiency $\eta(\lambda)$ is governed by a number of physical effects:

- Photon absorption coefficient as a function of wavelength λ
- Optical reflection losses, this was already treated in an earlier section
- Surface traps or recombination centres
- Photon generated charge carriers produced further from the depletion layer than a diffusion length

The spectral voltage responsivity, assuming an internal detector resistance of R_{pd} follows from:

$$R^V_{pd}(\lambda,\nu) = 0.808\lambda\eta(\lambda)R_{pd} \quad [\text{Volt Watt}^{-1}] \tag{7.86}$$

The above expressions are only valid up to the cut-off frequency ν_c of the photovoltaic device. In the next section, the frequency response of the photodiode will be addressed in some detail.
Since I_{ph} is the short circuit current, see figure (7.14), the definition of the current responsivity implicitly includes the notion that the device is working into a short circuit. This is commonly arranged by the usage of a trans-impedance amplifier that replaces the load impedance $\bar{Z}_L$, they are both displayed in figure (7.16) when dealing with photodiode equivalent circuits in the next paragraph.

7.3.2 Temporal frequency response

A photovoltaic device like a photodiode or a phototransistor has a frequency response that is determined by the characteristics of the device itself and by the circuitry in which it operates. Figure (7.16) shows at the left side the current equivalent circuit of a photodiode. From the equivalent circuit we can evaluate the total resistance to be:

$$R_T = \left(\frac{1}{R_{pd}} + \frac{1}{R_s + R_L}\right)^{-1} \quad \Rightarrow \text{cut-off frequency} \quad \nu_c = \frac{1}{2\pi R_T C_{pd}} \tag{7.87}$$

Let us take, as an example, the specific case of a Si-photodiode with an active sensor area for light detection of 1 mm^2 at a *reverse bias* of $V_{ex} = 4.3$ Volt, constituting a potential barrier of $\Phi = \Phi_{bi} + V_{ex} = 5$ Volt. Earlier, we derived an expression for the transition capacitance associated with the depletion region in a pn-junction, see equation (4.41). In working out a numerical example in (4.42), we obtained $C_T = 5.3 \cdot 10^{-4}$ F m^{-2} for a Si-junction with a *forward bias* of $V_{ex} = 0.5$ Volt, resulting in a potential barrier $\Phi = \Phi_{bi} - V_{ex} = 0.2$ Volt. Since the transition capacitance is proportional to $\Phi^{-1/2}$, the reverse bias chosen in the present example reduces C_T with a factor 5,

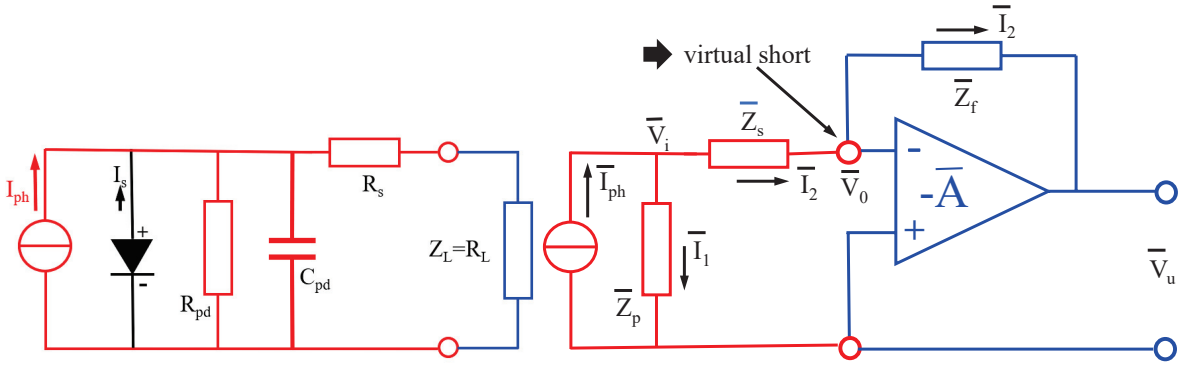


Figure 7.16: *The current equivalent circuit of a photodiode, it features the high but finite resistance R_{pd} of a reverse-biased pn-junction, the transition capacitance C_{pd} of the depletion layer, the small series resistance of the ohmic contact R_s and a load impedance $\bar{Z}_L$.*

owing to the much larger extent of the depletion region. In this way we arrive at a value $C_{pd} \approx 10^{-10}$ F for a sensor active area of 1 mm^2. Assuming for the resistors the following values: $R_d = 10^9\ \Omega$ and $R_s = 0.01\ \Omega$, a load resistance $R_L = 50\ \Omega$, we can essentially conclude that $R_T \approx R_L = 50\ \Omega$ from equation (7.87). This results in a cut-off frequency of $\nu_c = 30$ MHz.

If accessibility to higher frequencies is desirable, the junction capacitance can be further lowered by increasing the reverse bias voltage on the photodiode. If V_{ex} would be raised to say 50 Volt reverse bias, the cut-off frequency can be increased to ≈ 100 MHz. On the other hand, raising the reverse bias voltage level will increase the noise level of the device, so a careful trade-off between noise and frequency response needs to be made in practice. The junction capacitance can also be lowered by placing a thin intrinsic Si-layer in the pn-junction, a so-called *pin-photodiode*.

The right part of figure (7.16) shows the coupling of a photodiode to a trans-impedance amplifier that acts in very good approximation as a *virtual short* for the photocurrent generated by the device. Basically a transimpedance amplifier constitutes a current to voltage converter almost exclusively employing operational amplifiers (*op-amps*). Let us analyze in some detail the response characteristics of this frequently applied sensor interface.

In figure (7.16, right part) we have a parallel impedance $\bar{Z}_p$, comprising the pn-junction resistance R_{pd} parallel to a capacitive loading C_T, a coupling impedance $\bar{Z}_s$ and a feedback impedance $\bar{Z}_f$ in the amplifier that is characterized by a (complex) amplification parameter, the *open loop gain* $\bar{A}$. We have neglected the saturation current contribution of the pn-junction (by omitting the diode symbol) since in practice $\bar{I}_s \ll \bar{I}_{ph}$. Hence we have the following network equations, where it has been assumed that the resistive input impedance $\bar{Z}_i$ of the amplifier can be put as ∞, which is certainly admissible in case of a FET input stage:

$$\begin{aligned} \bar{I}_{ph} &= \bar{I}_1 + \bar{I}_2 \\ \bar{V}_i &= \bar{I}_1 \bar{Z}_p \\ \bar{V}_i - \bar{V}_0 &= \bar{I}_2 \bar{Z}_s \\ \bar{V}_0 - \bar{V}_u &= \bar{I}_2 \bar{Z}_f \\ \bar{V}_u &= -\bar{A}\,\bar{V}_0 \end{aligned}$$

Elimination of $\bar{I}_1, \bar{I}_2, \bar{V}_i$ and $\bar{V}_0$ from these equations yields:

$$-\bar{V}_u = \bar{I}_{ph}\bar{Z}_p \left[\left(\frac{1+\bar{A}}{\bar{A}} \right) \frac{\bar{Z}_s}{\bar{Z}_f} + \frac{1}{\bar{A}} + \left(\frac{1+\bar{A}}{\bar{A}} \right) \frac{\bar{Z}_p}{\bar{Z}_f} \right]^{-1} \tag{7.88}$$

In most practical cases the coupling impedance $\bar{Z}_s$ is quite small and can be disregarded, equation (7.88) then reduces to:

$$-\bar{V}_u = \bar{A}\ \bar{I}_{ph} \frac{\bar{Z}_p \bar{Z}_f/(1+\bar{A})}{\bar{Z}_p + \bar{Z}_f/(1+\bar{A})} \tag{7.89}$$

This expression is equivalent to a situation where the feedback impedance $\bar{Z}_f$ is replaced by an impedance $\bar{Z}_f/(1+\bar{A})$ parallel to $\bar{Z}_p$ at the input, constituting a very low impedance at the amplifier input and hence providing effectively a *virtual short* at $\bar{V}_0$ as was implicitly assumed in the definition of the spectral current responsivity above. We can rewrite equation (7.89) in terms of the complex feedback ratio $\bar{\beta}$:

$$-\bar{V}_u = \bar{I}_{ph} \frac{\bar{Z}_f}{1 + 1/(\bar{A}\,\bar{\beta})}, \qquad \text{with} \qquad \bar{\beta} = \frac{\bar{Z}_p}{\bar{Z}_p + \bar{Z}_f} \tag{7.90}$$

Let's consider the following cases:

- $|\bar{A}\,\bar{\beta}| \gg 1$

  Low frequency limit: $\Rightarrow -\bar{V}_u = \bar{I}_{ph}\bar{Z}_f$, full virtual short $Z_f/(1+\bar{A}) \Rightarrow 0$

- $|\bar{A}\,\bar{\beta}| \leq 1$
  As can be deduced from equation (7.90), the tipping point for the frequency response will be at $|\bar{A}\ \bar{\beta}| = 1$. To assess the position of this frequency in a bode diagram, we will treat the following representative example.
  We assume an op-amp comprising a *single pole* system $\bar{A} = A_0/(1+j\omega\tau_A)$, furthermore the capacitive loads at the op-amp input will allegedly be of major significance for the frequency characteristics of the system, these include the transition capacitance of the pn-junction associated with the photodiode C_{pd} parallel to the input capacitance of the op-amp C_{op}. Disregarding the (extremely high) internal resistance of the photodiode $R_{pd} \geq 1000\,M\Omega$, we designate the impedance $\bar{Z}_p = 1/j\omega C_T$ in figure (7.16) with $C_T = C_{pd} + C_{op}$, moreover we choose a resistive feedback $\bar{Z}_f = R_f$. This yields the following expression for the feedback ratio:

$$\overline{\beta} = \frac{1}{1 + j\omega R_f C_T} \quad \Rightarrow \quad \frac{1}{\bar{\beta}} = 1 + j\omega R_f C_T \tag{7.91}$$

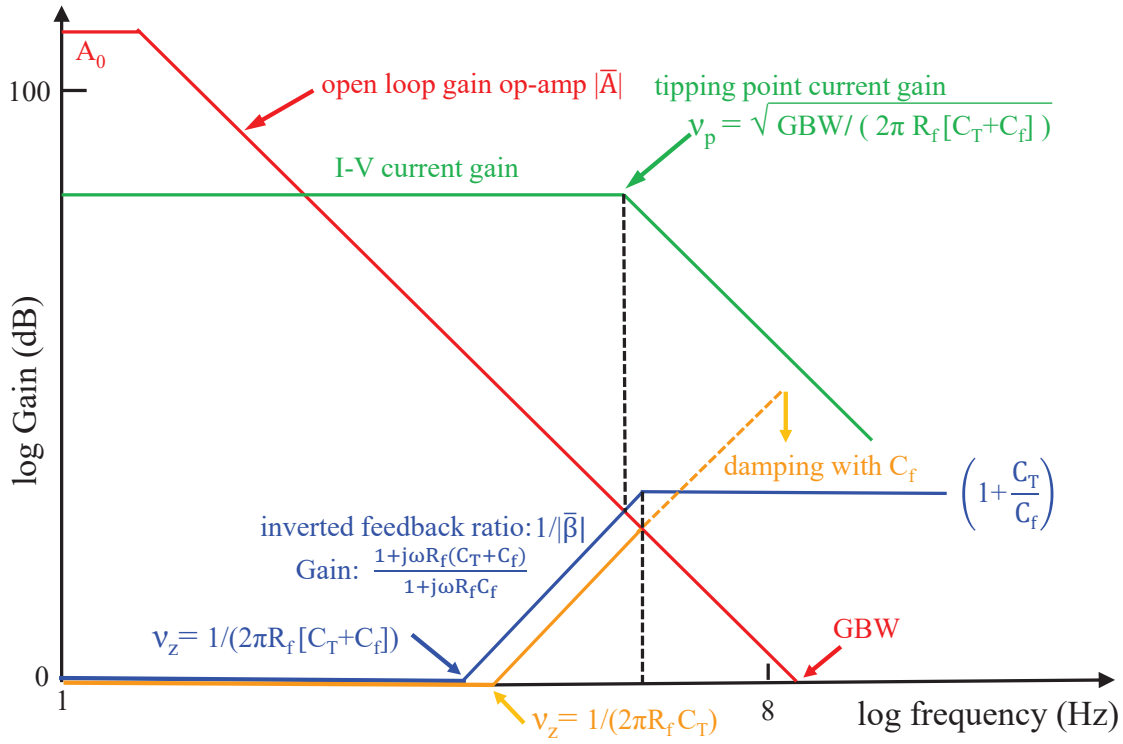


Figure 7.17: *Bode diagram of the trans-impedance amplifier read-out of a photodiode.*

The expression for $1/\bar{\beta}$ shows a *zero* at the tipping frequency $\nu = 1/(2\pi R_f C_T)$. The crossing with the single pole asymptote of the amplification $\bar{A}$, at which point we have $\bar{A}\ \bar{\beta} = 1$, would then in principle determine the tipping frequency of the system at $\nu = 1/(2\pi R_f C_T)$. However this set-up would render the feedback unstable, since the $\bar{A}$-pole and the $1/\bar{\beta}$-zero both introduce a *phase shift* of 90 degrees which, including the 180 degrees phase shift of the inverting amplifier, adds up to a full 360 degrees that introduces positive feedback and alleged oscillatory behavior. To circumvent this problem we can introduce a *phase lag* in the feedback by employing a small feedback capacitor C_f parallel to R_f. We then have the following transfer functions for $\bar{\beta}$ and $1/\bar{\beta}$:

$$\bar{\beta} = \frac{1 + j\omega R_f C_f}{1 + j\omega R_f (C_T + C_f)} \quad \Rightarrow \quad \frac{1}{\bar{\beta}} = \frac{1 + j\omega R_f (C_T + C_f)}{1 + j\omega R_f C_f} \tag{7.92}$$

Figure (7.17) shows the relevant asymptotes in the Bode diagram. The Gain-Band-Width (GBW)-product arrow indicates the maximum frequency range of the op-amp at an amplification of 0 dB. The (orange) asymptote originating from the zero point frequency at ν_z intersects the pole (red) asymptote of $|\bar{A}|$, determining the frequency at which we have $|\bar{A}\bar{\beta}| = 1$. Since both asymptotes are perpendicular they constitute a isosceles triangle (on a log-log scale) from which the value of the damping capacitor C_f can be derived: $C_f = \sqrt{C_T/2\pi R_f (GBW)}$. Substituting this value into the transfer function for $1/\bar{\beta}$ of equation (7.92), we can construct the blue plot in figure (7.17) that now contains a tipping point at

pole frequency $\nu_p(\text{blue}) = 1/2\pi R_f C_f$ with a horizontal asymptote towards high frequencies at an amplification of $(1 + C_T/C_f)$. The intersection of this blue plot with the (red) op-amp asymptote ($|\bar{A}\bar{\beta}| = 1$) marks the tipping point of the (green) I-V gain curve. It can be easily derived that this cut-off frequency equals:

$$\nu_c = \nu_p(\text{green}) = \sqrt{\frac{GBW}{2\pi R_f(C_T + C_f)}} \tag{7.93}$$

In this way the pole in the blue bode plot will lead to a 45 degree phase lead and, consequently, to a 45 degree phase margin for positive feedback. From equation (7.93) it is clear that the I-V tipping frequency is inversely proportional to the square root of the feedback resistor R_f. Hence, if the bandwidth of the photodiode is important, then the best approach would be to have a moderate transimpedance gain stage, followed by a broadband voltage gain stage. Assuming an op-amp GBW of 150 MHz, a $(C_T + C_f) = 100pF$ and $R_f = 10k\Omega$, we find an I-V gain tipping point frequency $\nu_p(\text{green}) \approx 5$ MHz.

7.3.3 Noise sources and spectral $\mathbf{D}^*(\lambda, \nu)$

From the photodiode equivalent circuit shown in figure (7.16) we can identify the potential noise contributers: shot noise from the incident photon flux density $\Pi_{ph}(\lambda)$, thermal noise from the detector resistance R_{pd} and the feedback resistance R_f, the inherent 1/f low-frequency-noise and the noise-equivalent-voltage at the input of the trans-impedance amplifier $V_{A_{rms}}$ that includes all the internal noise sources of the amplifier circuitry referred back to its input. As we pointed out earlier, the 1/f noise contribution can be made insignificant in most cases by proper component and material selection, moreover $V_{A_{rms}}$ can in general be made small owing to an application optimized low-noise design. Omitting these contributions we can write for the total noise voltage $V_{T_{rms}}$, assuming a modulation of the photon flux density with frequency ν and a noise equivalent bandwidth $\Delta\nu_c$:

$$V_{T_{rms}} = \left(2q^2\eta(\lambda)\Pi_{ph}(\lambda)R_{pd}^2 A_d \Delta\nu_c + 4kT_d R_{pd}\Delta\nu_c + 4kT_f R_f \Delta\nu_c\right)^{\frac{1}{2}} \tag{7.94}$$

Using this expression for the rms-noise-voltage, and employing the spectral voltage responsivity $R^V(\lambda, \nu)$, we can write:

$$\begin{aligned} R^V(\lambda,\nu) &= \frac{q\eta\lambda}{hc}R_{pd} \quad \Rightarrow \quad D^*(\lambda,\nu) = \frac{\sqrt{A_d\Delta\nu_c}}{V_{T_{rms}}}R^V(\lambda,\nu) \\ D^*(\lambda,\nu) &= \frac{\lambda}{hc}\sqrt{\frac{\eta(\lambda)}{2\Pi_{ph}(\lambda) + (4kT_{pd})/(q^2\eta R_{pd}A_d) + (4kT_f R_f)/(q^2\eta R_{pd}^2 A_d)}} \end{aligned} \tag{7.95}$$

where $\lambda \leq \lambda_{max}$.
Obviously the ultimate performance will be achieved when the noise due to the photon irradiance dominates. This implies:

$$\Pi_{ph}(\lambda) \gg \frac{2k}{q^2\eta R_{pd}A_d}\left(T_{pd} + T_f\frac{R_f}{R_{pd}}\right) \tag{7.96}$$

This condition may be fulfilled if the temperature of the sensor and the preamplifier are sufficiently low, the quantum efficiency for photon detection is high and the 'RA-product' is high. This RS-product should actually be regarded as one parameter, since R and A are correlated: when the sensitive area A goes up, the associated internal resistance R_{pd} normally decreases. So when the condition of equation (7.96) is met, the sensor can achieve radiation Background LImited Performance (BLIP). Substituting $\Pi_{ph}(\lambda) = B_{ph}(\lambda)$ we have:

$$D^*_{BLIP}(\lambda,\nu) = \frac{\lambda}{hc}\sqrt{\frac{\eta(\lambda)}{2B_{ph}(\lambda)}} \tag{7.97}$$

This expression is similar to the one we obtained for a BLIP-limited photoconductive device, the $D^*_{BLIP}(\lambda,\nu)$ for a photovoltaic device being superior by a factor $\sqrt{2}$. This originates from a $\sqrt{2}$-difference in noise level between GR-noise and shot noise as we showed previously in the treatment of noise sources.
When $B_{ph}(\lambda)$ becomes very low, like in space or astronomical applications, the condition (7.96) cannot hold and the sensor will become thermal-noise-limited. With an appropriate choice of R_f and T_f one might achieve a *sensor thermal-noise-limited* performance, in that case we have:

$$D^*_{th}(\lambda,\nu) = \frac{q\eta\lambda}{2hc}\sqrt{\frac{R_{pd}A_{pd}}{kT_{pd}}} = \frac{R^I(\lambda,\nu)}{2}\sqrt{\frac{R_{pd}A_{pd}}{kT_{pd}}} \tag{7.98}$$

Note that in all cases a high value of $R_{pd}A_{pd}$ is advantageous.

7.4 Solar cell

7.4.1 Operation principle

Construction wise the solar cell shows a large similarity to the photodiode, however in this case no external voltage is applied to the pn-junction. The solar cell itself generates electric power and, consequently, fits in the category of self-generating transducers. The sensor constitutes a pn-junction that is exposed to light. Figure (7.18) shows the band structures for two different situations: *unexposed* and *exposed* to light.
In the situation without illumination the pn-junction is characterized by a relatively thin depletion layer since the external voltage bias is absent ($W \propto \Phi_{bi}$) contrary to the situation of the photodiode that is subjected to a large reverse bias. When exposed to light a photocurrent is generated that flows as a drift current of minority charge carriers through the depletion layer. Since there is no external voltage bias, this drift current needs to be compensated by a reverse diffusion current flow of majority charge carriers induced by a shift towards forward bias of the pn-junction. Far away from the depletion layer at approximately four diffusion lengths (either L_n or L_p) this causes a difference in Fermi potential between the n-type and p-type regions, see Figure (7.18) lower right. This potential difference manifests itself as an open-circuit-voltage V_{oc}:

$$V_{oc} = (E_F/q)_n - (E_F/q)_p \tag{7.99}$$

The value of V_{oc} can be derived from the condition for current compensation:

$$\begin{aligned} I_{dif} &= I_{ph} \Rightarrow I_{dif} = I_s \left[e^{(qV_{oc}/kT)} - 1\right] \Rightarrow \\ V_{oc} &= \frac{kT}{q} ln(1 + I_{ph}/I_s) \approx \frac{kT}{q} ln\left(\frac{j_{ph}}{j_s}\right) \end{aligned} \tag{7.100}$$

where we have removed the dependence on the solar cell area by switching to the current densities j_{ph} and j_s and where the approximation is justified because of $j_{ph} \gg js$.
It is clear from expression (7.100) that to maximize V_{oc} it is important to have, apart from a high j_{ph}, the lowest possible value for the leakage current density j_s. The leakage current density can be expressed in terms of the equilibrium concentrations of the minority charge carriers p_{ne} and n_{pe} in the n-type and p-type material respectively, their diffusion lengths L_p, L_n and their life times τ_p, τ_n. We can write:

$$j_s = q\left(\frac{p_{ne}L_p}{\tau_p} + \frac{n_{pe}L_n}{\tau_n}\right) \quad \Rightarrow \quad j_s = q\left(\frac{p_{ne}D_p}{L_p} + \frac{n_{pe}D_n}{L_n}\right) \tag{7.101}$$

with the substitution $L_p = (D_p\tau_p)^{1/2}, L_n = (D_n\tau_n)^{1/2}$.
Making use of the relations derived earlier for extrinsic semiconductors $p_{ne} = n_i^2/N_d$ and

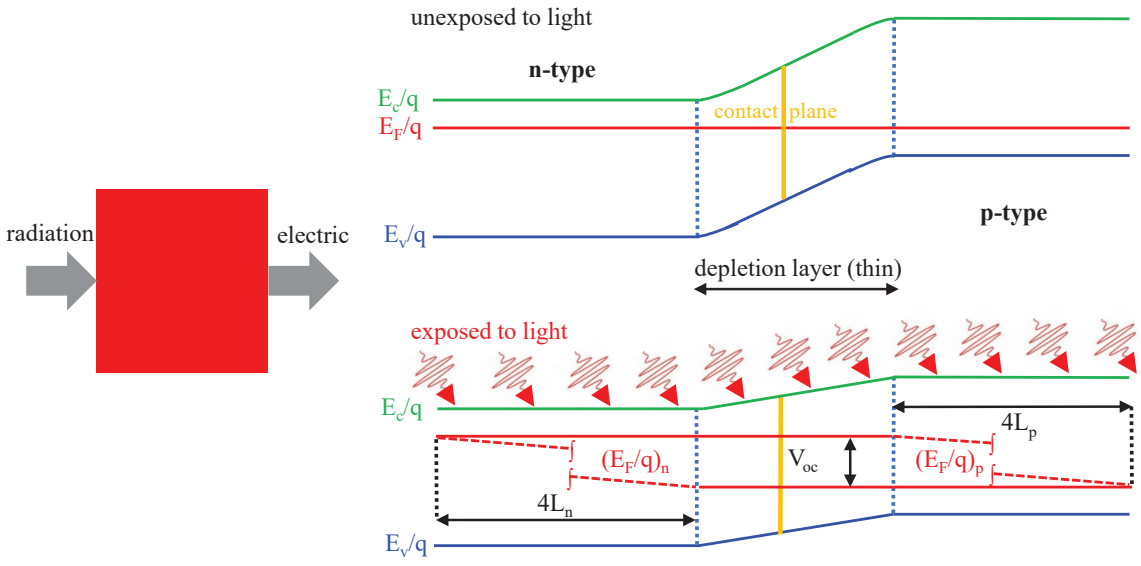


Figure 7.18: *Band diagram of a solar cell. Since no external bias is applied, the depletion layer is a lot thinner than in the case of the photodiode that operates under substantial reverse bias.The charge carriers contributing to the photocurrent originate predominantly in the region adjacent to the depletion layer with an extent of approximately four diffusion lengths (either L_p or L_n). The difference in Fermi potential between the n-type and p-type regions at the termination of the contributing diffusion regions constitutes the open-circuit voltage V_{oc}, with oc short for 'open-circuit'.*

$n_{pe} = n_i^2/N_a$ we arrive at an expression for j_s as a function of the impurity concentrations of the p-type and n-type material:

$$j_s = qn_i^2 \left(\frac{D_p}{N_d L_p} + \frac{D_n}{N_a L_n} \right) \tag{7.102}$$

In order to obtain a low value of j_s, a high concentration of impurities is apparently favorable. For silicon as base material this will work up to impurity concentrations of $N \approx 10^{24}$ m^{-3}, higher values will not add anymore to the lowering of I_s, since in that case the diffusion length L will decrease inversely with N, ($L \propto N^{-1}$) due to Auger recombination effects. As a consequence the leakage current I_s will then remain almost constant. So V_{oc} is a measure of recombination in the device. Laboratory crystalline silicon solar cells have V_{oc} of up to 720 mV, while commercial solar cells have V_{oc} exceeding 600 mV.

7.4.2 Equivalent circuit and maximum deliverable power

The current density-voltage (j-V) characteristic that describes the operation of an illuminated solar cell that behaves as an ideal diode is given by:

$$j(V) = j_d - j_{ph} = j_s \left(e^{(qV/kT)} - 1 \right) - j_{ph} \tag{7.103}$$

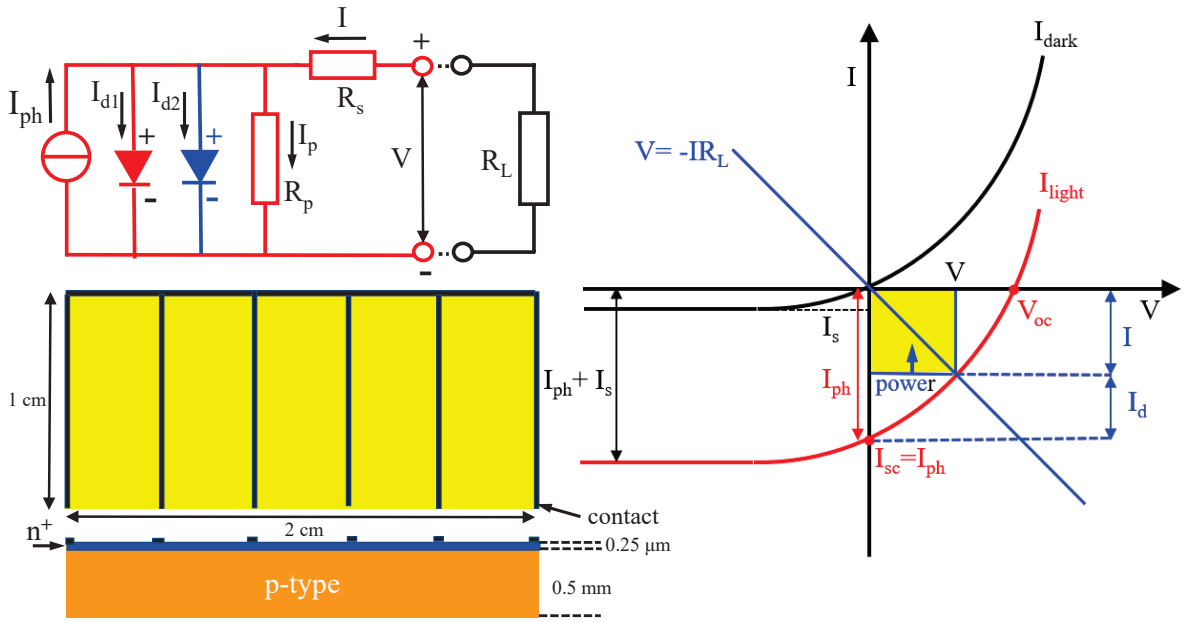


Figure 7.19: *Upper left: one-diode equivalent circuit of a solar cell (red) and two-diode equivalent circuit (blue addition). Current equation:* $I = I_{d1} + I_{d2} + I_p - I_{ph}$*, with* I_{d1} *and* I_{d2} *the compensating forward diffusion currents (majority charge carriers) and* I_{ph} *the reverse photocurrent (minority charge carriers). Right part: Power characteristic of a solar cell for an ideal diode. The yellow area designates the deliverable power into a load resistance* R_L*.*

The equivalent circuit can simply be described as a current source and a diode that are connected in parallel. In practice, however, the characteristic is influenced by the presence of a series resistance R_s (contacts and leads) and a shunt resistance R_p (manufacturing imperfections and defects). The influence of this parasitic resistivity can be addressed with the aid of the equivalent circuit shown in the left part of figure (7.19). Let's first deal however with the case of ideal diode behavior
In that case the available power can be derived from the I-V characteristic for an exposed solar cell, see the diagram in the right part of figure (7.19). The yellow area, defined by the intersection of the load line $V = -IR_L$ with the cell's I-V characteristic represents the deliverable power for that particular value of the resistive load. When a load is directly connected to a solar cell, the operating point will rarely be at peak power. By varying the impedance seen by the solar cell, the operating point can be moved towards the peak power point. For the position of the maximum power point the condition $dP/dV = 0$ should hold, with $P = I(V) \cdot V$. So we have:

$$\frac{dP}{dV} = V\frac{dI}{dV} + I(V) = 0 \quad \Rightarrow \quad \frac{dI}{dV} = -\frac{I(V)}{V} \tag{7.104}$$

Hence, the maximum power point is achieved when the incremental conductance is equal to the negative of the instantaneous conductance. If we designate the maximum power $P_{max} = I_{mp}V_{mp}$, (mp for *maximum power*) and divide P_{max} by the product $I_{sc}V_{oc}$, we obtain the power filling factor:

$$f_P = \frac{I_{mp}V_{mp}}{I_{sc}V_{oc}} \tag{7.105}$$

The filling factor f_P does not change drastically with the specific value of the V_{oc}, a commercial Si-cell with $V_{oc} = 600mV$ entails a $f_P \approx 0.83$, a laboratory Si-cell with $V_{oc} = 720mV$ a $f_P \approx 0.85$. A different material like for example GaAs can approach a $f_P \approx 0.90$.

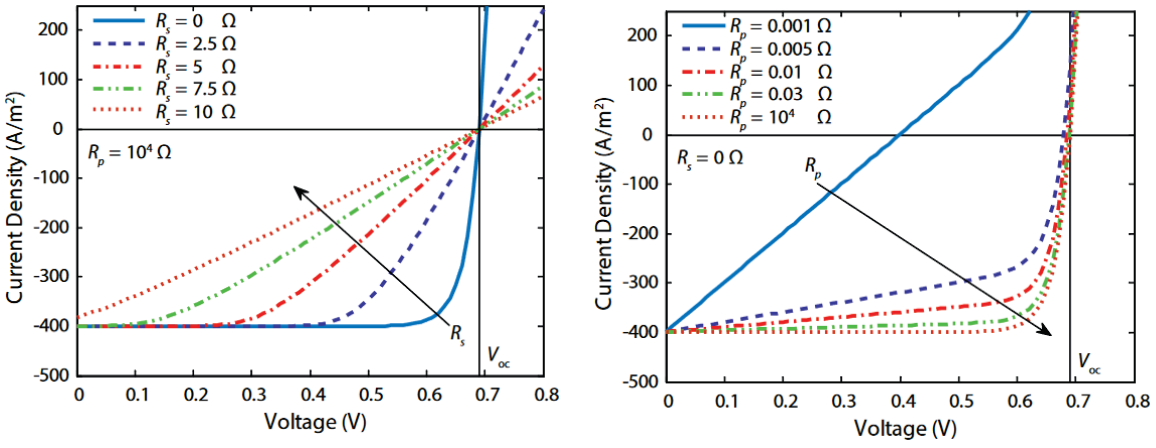


Figure 7.20: *Current density-Voltage (j-V) characteristics with parasitic series, R_s, and parallel, R_p resistor values as parameters.*

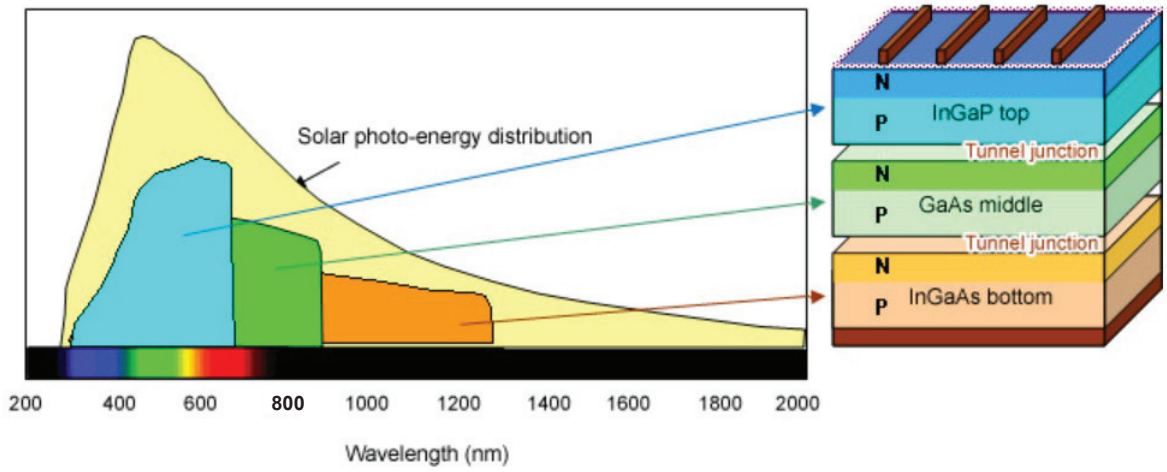


Figure 7.21: *Coverage of the standard solar photosphere spectrum with a tri-layer junction stack comprising Indium- Gallium-Phosphide (InGaP) with the shortest wavelength cut-off, a middle layer of Gallium Arsenide (GaAs) and a bottom junction of Indium Gallium Arsenide (InGaAs) that reaches into the infrared region.*

In addition we can define the *characteristic resistance* $R_{mp} = V_{mp}/I_{mp}$ of a solar cell. This entails the cell's output resistance at its maximum power point. If the resistance of the load is equal to the characteristic resistance of the solar cell, then the maximum power is transferred to the load, and the solar cell operates at its maximum power point. In practice a good estimate can be obtained from the ratio V_{oc}/I_{sc} being $\approx V_{mp}/I_{mp}$. Commercial silicon solar cells are very high current and low voltage devices. A 156 mm (6 inch) square solar cell has a current of almost 9 amps and a maximum power point voltage of 0.6 volts giving a characteristic resistance, R_{mp}, of 0.067 Ω. A 72 cell module from the same cells has $R_{mp} = 4.8\ \Omega$. A lead series resistance of 30 mΩ has a negligible effect on a full module but has a catastrophic effect on a single cell coupon. Thus the characteristic resistance is also useful because it puts series and shunt resistance in context. Referring to the equivalent circuit in figure (7.19) we can accommodate the influence of R_s and R_p on the voltage dependent current density $j(V)$:

$$j(V) = j_s \left(e^{\frac{q(V - Aj(V)R_s)}{kT}} - 1 \right) + \frac{V - Aj(V)R_s}{R_p} - j_{ph} \quad j(V), j_{ph} \text{ reverse!} \tag{7.106}$$

where A is the area of the solar cell. The effect of R_s and R_p as parameters on the j-V characteristic is shown in figure (7.20). These effects will of course also impact the achievable filling factor f_P. In practical solar cells the saturation current does not really follow the Boltzmann approximation, to compensate for this a so-called *ideality factor* n is introduced in the Boltzmann law: $e^{qV/nkT}$ to better fit the experimental data. This non-ideal behavior can be incorporated in a two-diode equivalent circuit, the j-V

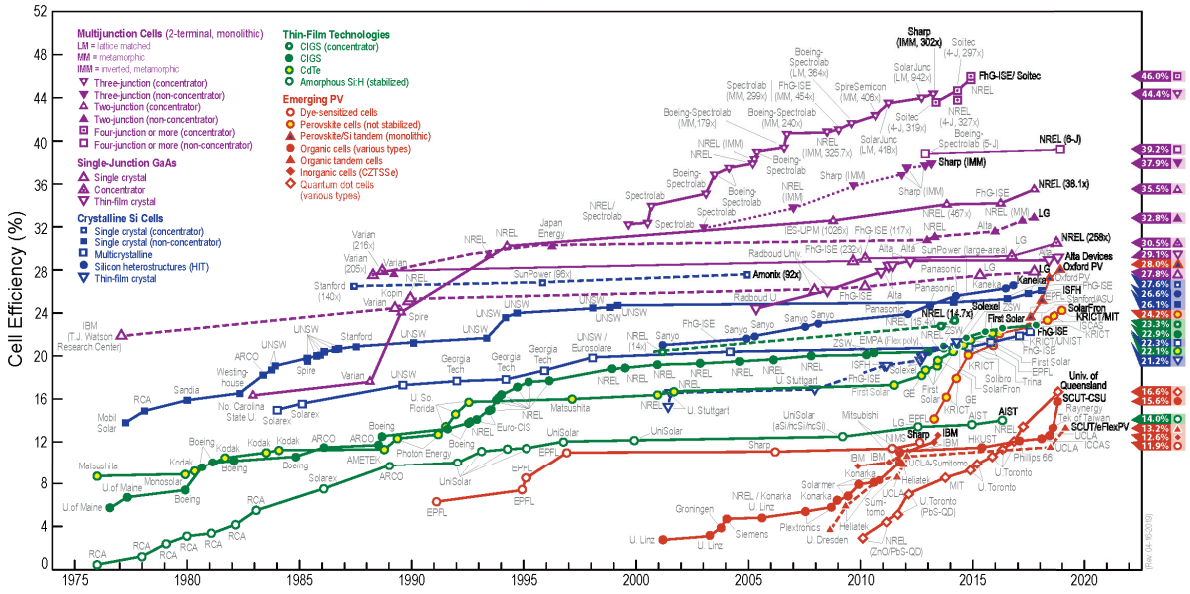


Figure 7.22: *Evolution over the years of power conversion efficiency of solar cells in researching several technologies like single/multijunction cells, thin film technology cells, metal-organic compound cells etc.*

characteristic is then given by:

$$j(V) = j_{s1}\left(e^{\frac{q(V-Aj(V)R_s)}{n_1 kT}} - 1\right) + j_{s2}\left(e^{\frac{q(V-Aj(V)R_s)}{n_2 kT}} - 1\right) + \frac{V - Aj(V)R_s}{R_p} - j_{ph} \quad (7.107)$$

A typical lay-out of a crystalline silicon solar cell is also sketched in figure (7.19) It consists of a thick p-type substrate with $N_a \approx 10^{22}$ m^{-3} (1 Ω·cm) with on top a thin layer of heavily doped n-type material. The ohmic contact on the n-material has a comb like structure to keep the series resistance of the cell at an absolute minimum to avoid power efficiency loss.
Since wavelengths with $\lambda \geq \lambda_{max}$ will not interact with a semiconductor material, it remains transparent and another layer with a higher cut-off wavelength could be applied to raise the efficiency over the solar spectrum. This has led to the development of sandwich layer solar cells. An example is shown in figure (7.21) comprising three compound semiconductors with both n-type and p-type impurities: an InGaP/GaAs/InGaAs multijunction cell. Another example in this category is CdS-CdSe-CdTe solar battery. Both tri-layers reach conversion efficiencies between 0.35-0.40. Many new design concepts have been developed over the years including thin film technology and the application of metal-organic compounds. These developments in research on photovoltaic devices are summarized in figure (7.22).

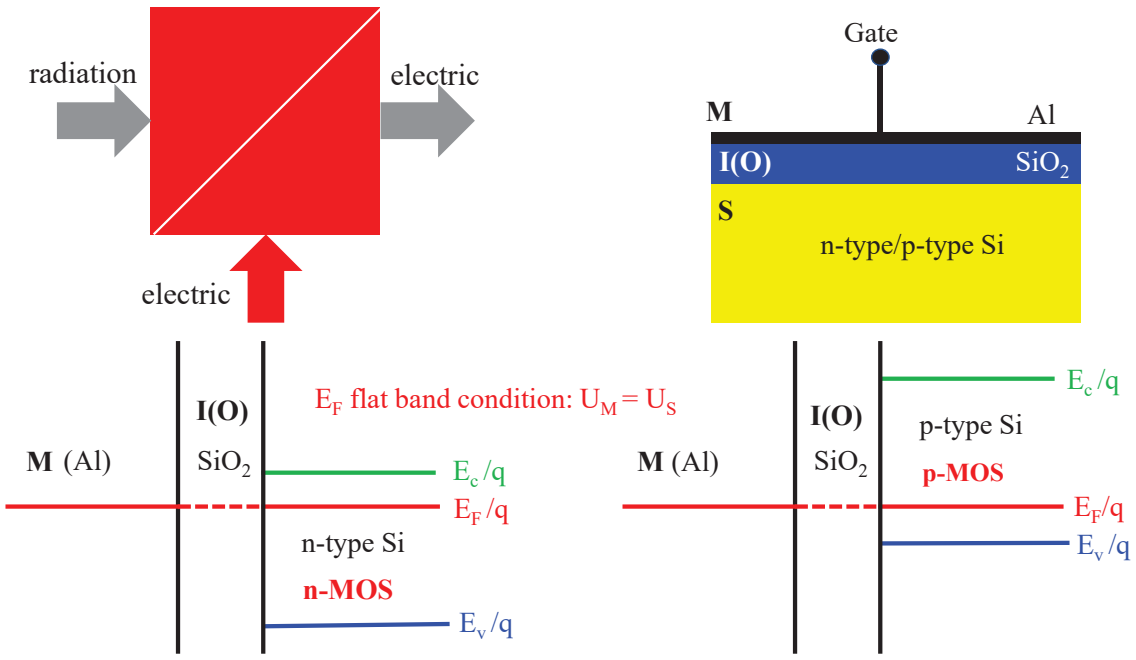


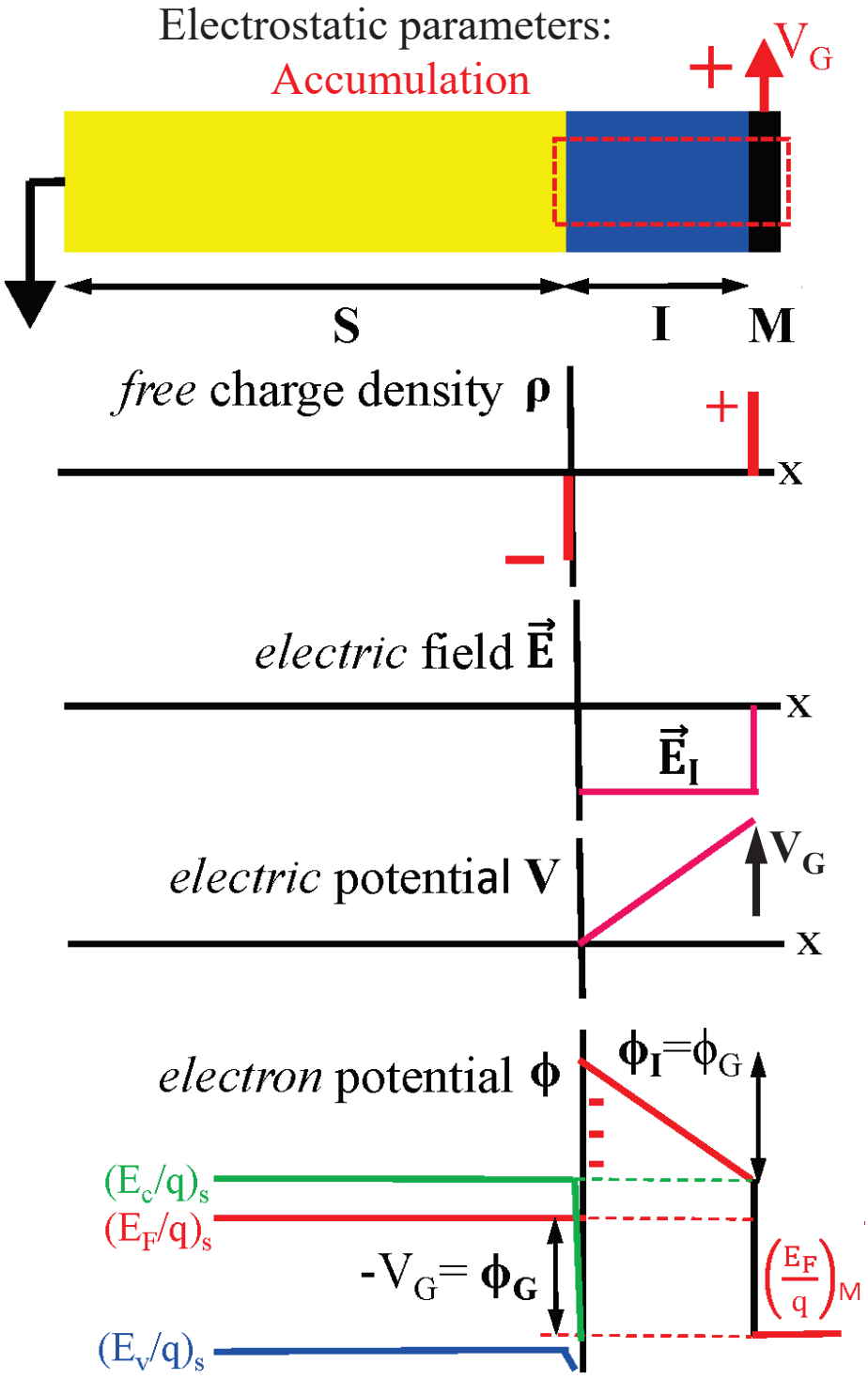
Figure 7.23: *MIS (MOS) configuration, a modulated transducer. The flat band condition implies equalization of the Fermi potential levels between the metal electrode (gate) and the semiconductor, that are separated by an insulating layer (mostly an Oxide).*

7.5 MIS/MOS element

7.5.1 Operation principle

A light sensitive MIS (Metal-Insulator-Semiconductor) diode is applied as a pixel image element in solid state picture devices and belongs to the category of modulating transducers. A schematic is given in figure (7.23). Since in most cases the insulating layer comprises an oxide, the MIS element is also referred to as a MOS diode.
In thermal equilibrium the band diagram for n-type or p-type silicon is also displayed in figure (7.23). The situation shown implicitly assumes that the MIS meets the so-called 'flat-band' condition: equalization of the Fermi potential levels between the metal and the semiconductor implying that the exit potentials U_M and U_S have the same value. In most cases a fixed surface charge is present at the interface between the insulator and the semiconductor owing to ionized impurities or crystal imperfections at the semiconductor surface. Moreover, initially charge effects are also induced by a non-zero difference between U_M and U_S that works towards equalization. Therefore we have assumed that the flat-band condition has been met in figure (7.23) by virtue of fixed surface charges at the I-S interface. More general however: if the MIS contains a particular fixed surface charge density Q_f, a certain external bias voltage V_{FB} might be necessary to meet the flat-band condition, the so-called flat-band voltage. In the treatment of the MIS physics that follows it has been assumed that the flat-band condition has been satisfied.
In describing the physical behavior of a MIS-element three cases can be distinguished.

The example we shall analyze here entails an n-type MIS element where the semiconductor is grounded. In that case we have the following possibilities:
- Accumulation $V_{gate} > 0$
- Depletion $V_{gate} < 0$
- Inversion $V_{gate} \ll 0$



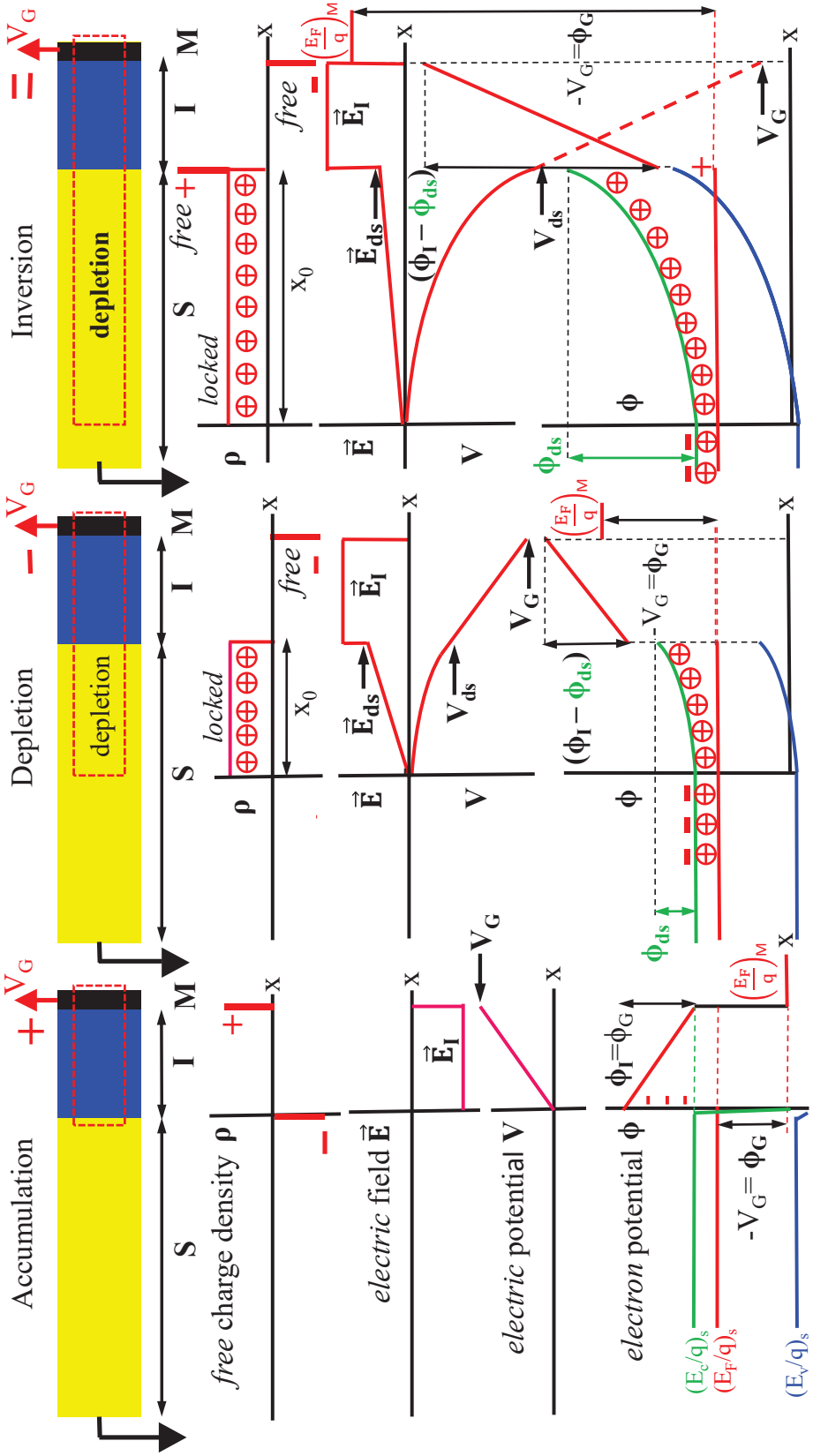


Figure 7.24: *Electrostatics of Accumulation, Depletion and Inversion in a MIS element.*

Let's now explore the electrostatic configurations of the MIS element under these three biasing conditions.

- Accumulation
  The charge distribution, the electrical field strength, the electrical potential and the band diagram in *electron*-potential are shown in figure (7.24) in the left panel. The positive voltage on the gate electrode causes that free electrons present at the gate-insulator interface are pushed back through the attractive force exerted by the positive bias voltage. This gives rise to the creation of a positive surface charge of free charge carriers at the metal-insulator interface. This positive surface charge induces a negative surface charge at the semiconductor-insulator interface fed by the majority charge carriers (electrons) that are abundantly present in the n-type semiconductor. As a consequence the electric field $\vec{E}$ associated with these surface charges exhibits a step behavior at both the metal and the semiconductor side of the insulator: within the insulating layer we have a constant field strength $\vec{E_I}$, outside the field strength is zero. The presence of a negative surface charge of free electrons at the boundary of the semiconductor manifests itself as a steep downward decline of the conduction and the valence band electron potentials: there exists a deep potential well at this interface plane for electrons ('accumulation') and, conversely, a potential maximum for holes, the presence of which is consequently excluded at this surface.

- Depletion
  The charge distribution, the electrical field strength, the electrical potential and the band diagram in *electron*-potential are shown in figure (7.24) in the middle panel. At the metal-insulator surface we now have a negative surface charge and by induction the electrons in the semiconductor will be repelled at the semiconductor-insulator surface, this process results in the formation of a depletion layer of positively charged donor atoms in the n-type silicon. These positive charges are locked in the crystal lattice and constitute a space charge with the density of the doping concentration N_d, resulting in a depletion layer with a thickness x_0, see figure (7.24).
  - *Charge distribution*
  The charge per unit area at the interface plane of the semiconductor and the insulator amounts to $Q_D = qN_dx_0$. Consider a closed surface, marked by the dotted line in figure (7.24), with end-faces just outside the depletion layer and just inside the metal electrode. Outside these end-face boundaries we obviously have $\vec{E} = 0$. Applying Gauss's theorem in electrostatics we have:

$$\begin{aligned} \oiint_S \vec{D}\cdot\vec{n}\,dS &= \iiint_V \rho\,dV, \qquad \oiint_S \vec{D}\cdot\vec{n}\,dS = 0 \Rightarrow \iiint_V \rho\,dV = 0 \\ \Rightarrow Q_G + Q_D &= 0 \Rightarrow Q_G = -qN_dx_0 \end{aligned} \tag{7.108}$$

  where Q_G is the surface charge density on the metal electrode. The surface charge density Q_D constitutes a bound charge locked into the crystal lattice, whereas Q_G

represents a free charge.
- *Electric field*
In the depletion layer (depth x_0) applies:

$$\vec{E}_d = \int_0^x \rho \, d\vec{x} = \frac{qN_d\vec{x}}{\epsilon_0\epsilon_s}, \text{ at } |\vec{x}| = x_0 \Rightarrow |\vec{E}_{ds}| = \frac{qN_dx_0}{\epsilon_0\epsilon_s} \tag{7.109}$$

At the boundary plane with the insulator applies (with $\vec{n}$ the unit vector normal to the interface plane):

$$\vec{D}_{ds} \cdot \vec{n} = \vec{D}_I \cdot \vec{n}, \;\; \vec{n} \text{ parallel to } \vec{D} \;\; \Rightarrow |\vec{D}_{ds}| = |\vec{D}_I| \Rightarrow \frac{|\vec{E}_I|}{|\vec{E}_{ds}|} = \frac{\epsilon_s}{\epsilon_I} \tag{7.110}$$

Hence the discrete jump in field strength at the interface between the semiconductor and the insulator is determined by the ratio of their respective dielectric constants. $\vec{E}_I$ remains constant in the insulator when no built-in residual charge is present.
- *Potential distribution and band diagram*
The potential in the depletion layer follows from:

$$V_d(x) = -\int_0^x \vec{E}_d \cdot \, d\vec{x} = -\frac{qN_dx^2}{2\epsilon_0\epsilon_s}, \text{ at } x = x_0 \Rightarrow V_{ds} = -\phi_{ds} = -\frac{qN_dx_0^2}{2\epsilon_0\epsilon_s}, \tag{7.111}$$

with V_{ds} the electric potential at the boundary layer with the insulator and ϕ the corresponding electron potential. The potential in the insulator can be calculated from:

$$V_I(x) = -\frac{qN_dx_0^2}{2\epsilon_0\epsilon_s} - \int_{x_0}^x \vec{E}_I \cdot \, d\vec{x} = -\left[\frac{qN_dx_0^2}{2\epsilon_0\epsilon_s} + \frac{qN_dx_0}{\epsilon_0\epsilon_I}(x - x_0)\right] = -\phi_I(x) \tag{7.112}$$

At the interface plane between the insulator (thickness d) and the metal electrode we have $V_I(d) = V_G$, the externally applied bias voltage, for which the following relation holds:

$$V_G = -\phi_G = -\left(\frac{qN_dx_0^2}{2\epsilon_0\epsilon_s} + \frac{qN_dx_0}{\epsilon_0\epsilon_I}d\right) \tag{7.113}$$

The value of x_0, the depth of the depletion layer, is a function of the surface potential V_{ds} (see also the treatment of the pn-junction in Chapter 4):

$$x_0 = \left[-\frac{2\epsilon_0\epsilon_sV_{ds}}{qN_d}\right]^{\frac{1}{2}} \;\Rightarrow\; V_G = V_{ds} - \frac{\epsilon_s}{\epsilon_I}d\left[\frac{-2qN_dV_{ds}}{\epsilon_0\epsilon_s}\right]^{\frac{1}{2}} \tag{7.114}$$

The right hand term represents the voltage drop ($V_G - V_{ds}$) over the insulator, for a given specific value of V_{ds}. However in practice the gate voltage V_G is a

given, so we need to work out an explicit expression for the surface potential of the semiconductor V_{ds}. Rewriting equation(7.114) we have:

$$(V_G - V_{ds})^2 = (-2q\epsilon_0\epsilon_s N_d)\left(\frac{d}{\epsilon_0\epsilon_I}\right)^2 V_{ds} \tag{7.115}$$

Substituting the capacitance per unit area of the oxide layer $C_{ox} = \epsilon_0\epsilon_I/d$:

$$(V_G - V_{ds})^2 = \left(\frac{-2q\epsilon_0\epsilon_s N_d}{C_{ox}^2}\right) V_{ds} \tag{7.116}$$

Let us now introduce the quantity $V_0 = (q\epsilon_0\epsilon_s N_d)/C_{ox}^2$. V_0 represents a material constant with the dimension of a potential that depends on the capacitance per unit area of the insulating oxide, hence we get:

$$(V_G - V_{ds})^2 = -2V_0V_{ds} \Rightarrow V_G,\ V_{ds} \text{ are negative potentials, } V_0 \text{ is positive!} \tag{7.117}$$

Solving for V_{ds}, fulfilling the boundary condition $|V_{ds}| < |V_G|$:

$$V_{ds} = V_G - V_0 + (-2V_0V_G + V_0^2)^{1/2} \tag{7.118}$$

If we have a $V_0 = 0.3$ Volt and a gate voltage $V_G = -5$ Volt, expression (7.118) yields a potential V_{ds} at the insulator/semiconductor interface of ≈ -3.5 Volt. For completeness we should also derive a similar expression for the case that the MIS cell contains a p-type semiconductor. In that situation V_G constitutes a positive bias voltage, V_{ds} and V_{ox} are both also positive potentials according to:

$$V_G = V_{ds} + V_{ox}, \quad Q_G = -(-qN_a x_0) = +qN_a x_0 = -Q_D \tag{7.119}$$

As before we can now write:

$$(V_G - V_{ds})^2 = \left(\frac{2q\epsilon_0\epsilon_s N_d}{C_{ox}^2}\right) V_{ds} \tag{7.120}$$

Substituting V_0:

$$(V_G - V_{ds})^2 = 2V_0V_{ds}, \Rightarrow V_G, V_{ds} \text{ and } V_0 \text{ are now all positive potentials!} \tag{7.121}$$

Solving for V_{ds}, abiding the boundary condition $|V_{ds}| < |V_G|$:

$$V_{ds} = V_G + V_0 - (2V_0V_G + V_0^2)^{1/2} \tag{7.122}$$

- Inversion
  The charge distribution, the electric field strength, the electric potentials and the electron potentials are displayed in the right panel of figure (7.24). If the the applied negative bias voltage V_G exceeds a certain value, the conduction band and the valence band becomes so strongly bent that the Fermi level gets positioned inside the valance band, see also figure (7.24) right panel. At the interface plane between the semiconductor and the insulator this gives rise to a deep potential

minimum for holes and a potential maximum for electrons (see the graph for the electron potentials). As a consequence almost no free electrons can exist near the the semiconductor contact plane, which implies that the recombination probability for holes is greatly reduced. Hence at the boundary layer hole conduction will occur, i.e. conduction by virtue of the *minority charge carriers*. This hole conducting layer is designated the *inversion layer*, since the conduction by free charge carriers is not accommodated by majority charge carriers (electrons in the case of an n-type semiconductor) but by the minority charge carriers (holes in this case).

-*Charge distribution*

Applying again Gauss's theorem we get in this case:

$$Q_G + Q_D + Q_{inv} = 0 \quad \Rightarrow \quad Q_G = -(qN_d x_0 + Q_{inv}), \tag{7.123}$$

where Q_D is the bound, lattice locked, charge in the semiconductor, Q_G and Q_{inv} represent the negative and positive free charge respectively.

-*Electric field*

At the boundary plane with the semiconductor we now have:

$$\begin{aligned} \vec{D}_I \cdot \vec{n} &= \vec{D}_{ds} \cdot \vec{n} + Q_{inv}, \;\; \vec{n} \text{ parallel to } \vec{D} \;\; \Rightarrow |\vec{D}_I| = |\vec{D}_{ds}| + Q_{inv} \Rightarrow \\ |\vec{E}_I| &= \frac{\epsilon_s}{\epsilon_I}|\vec{E}_{ds}| + \frac{Q_{inv}}{\epsilon_s \epsilon_I} \quad \Rightarrow \\ |\vec{E}_{ds}| &= \frac{qN_d x_0}{\epsilon_0 \epsilon_s} \qquad \Rightarrow \qquad |\vec{E}_I| = \frac{qN_d x_0 + Q_{inv}}{\epsilon_0 \epsilon_I} \end{aligned} \tag{7.124}$$

Thus the discrete step in the field strength becomes now enlarged with an amount $Q_{inv}/\epsilon_0\epsilon_I$ as compared to the previous case where the inversion layer did not yet emerge.

-*Potential distribution and band diagram*

The gate voltage can now be written as:

$$V_G = -\phi_G = -\left(\frac{qN_d x_0^2}{2\epsilon_0 \epsilon_s} + \frac{(qN_d x_0 + Q_{inv})}{C_{ox}}\right) \tag{7.125}$$

The electrical potential at the surface of the semiconductor layer can then be expressed as:

$$V_{ds} = \left(V_G + \frac{Q_{inv}}{C_{ox}}\right) - V_0 + \left[-2V_0\left(V_G + \frac{Q_{inv}}{C_{ox}}\right) + V_0^2\right]^{\frac{1}{2}}, \tag{7.126}$$

with $V_0 = (\epsilon_0 \epsilon_s q N_d)/C_{ox}^2$ and $C_{ox} = \epsilon_0 \epsilon_I / d$ the capacitance per unit area of the insulating oxide layer with thickness d. Potentials V_G and V_{ds} are negative, whereas V_0 and Q_{inv}/C_{ox} represent positive potentials, the latter since it derives from hole storage in the inversion layer.

Similarly, as in the case of depletion, we can also give an expression for the case that the MIS cell is based on a p-type semiconductor. The relevant equation for V_{ds} in that case is:

$$V_{ds} = \left(V_G + \frac{Q_{inv}}{C_{ox}}\right) + V_0 - \left[2V_0\left(V_G + \frac{Q_{inv}}{C_{ox}}\right) + V_0^2\right]^{\frac{1}{2}}, \tag{7.127}$$

Potentials V_G, V_{ds} and V_0 are positive, whereas Q_{inv}/C_{ox} represents a negative potential since it derives from electron storage in the inversion layer. From these formula's we can also immediately derive the value of the flat-band voltage V_{fb} should the flat-band condition initially not be satisfied. If the spurious surface charge density and the effect of the difference between the workfunctions $U_M - U_S$ is combined in a certain charge density value Q_F, we apparently have, except for Q_{inv}, yet another surface charge component Q_F present. From the formula for V_G it becomes clear that the associated voltage amounts to Q_F/C_{ox}. This yields directly $V_{fb} = Q_F/C_{ox}$, indicating that the voltage V_G needs correction by V_{fb}. Hence we have for V_{ds}:

$$V_{ds} = (V_G - V_{fb}) + \frac{Q_{inv}}{C_{ox}} - V_0 + \left[-2V_0\left((V_G - V_{fb}) + \frac{Q_{inv}}{C_{ox}}\right) + V_0^2\right]^{1/2} \quad (7.128)$$

7.5.2 Frequency response: capacitance of a MIS element.

Since the behavior of a MIS element depends on the applied bias voltage, the small-signal capacitance $C_{ss} = dQ/dV$ will be different in each operational mode:
In case of *accumulation* we have two free surface charges at either side of the insulating layer:

$$C_{ss} = \frac{dQ}{dV} = A\,C_{ox},\ A = \text{MIS-area},\ C_{ox} = \text{capacitance/unit area of the oxide} \quad (7.129)$$

In case of *depletion* we encounter the oxide capacitance in series with the capacitance of the depletion region:

$$C_{ss} = \left(\frac{1}{C_{ox}} + \frac{1}{C_{dep}}\right)^{-1} \quad (7.130)$$

In the *inversion* mode we can discriminate between two frequency domains:
(i) a slow (low frequency) variation of the gate voltage V_G that give rise to a change of the free surface charge in the inversion layer and leaves the space charge locked in the depletion layer unimpaired, hence we have a capacitance $C_{ss} = C_{ox}$
(ii)a fast (high frequency) variation of the gate voltage V_G that leaves insufficient time for the inversion layer to follow or to even form by optical or thermal excitation. In that case the capacitance remains the series value shown in equation (7.130). In conclusion: in the inversion mode the MIS capacitance depends on the frequency of the bias voltage V_G.
The MIS structure is a fundamental building block of light sensitive Charged Coupled Devices applied in electronic camera's, the incident light then governs the charge content of the inversion layer.

7.6 Photomultiplier tube

Photomultiplier tubes belong to the category of modulating transducers. The most relevant properties can be summarized as follows:

- Large sensitive area per phototube up to five inch (12.5 cm) diameter.
- Very large sensitivity allowing measurement of extremely low light levels, including the ability of single photon counting.
- Fast temporal response, nano- to sub-nanosecond response times.
- Large dynamic range in wavelength, from near-IR to mid-UV.
- Large range of applications: optical spectroscopy, laser beam measurement, astronomical instrumentation, energetic radiation detection like relativistic particles, X-ray and gamma-ray radiation by measuring the scintillation light that is produced in a sensor comprising a gas-filled cell or an (an)organic crystal assembly.

A photomultiplier tube structure comprises two main elements:
- A photosensitive layer: *the photocathode*. This photocathode converts, with a certain quantum efficiency, the incident photons in low-energy photo-electrons that get emitted into vacuum by the *external* photo-electric effect.
- An *electron multiplier*. Since the light levels are often as low as a few to a few hundred photons, electron multiplier stages are added that yield charge levels of the order of 10^7-10^{10} electrons at the anode output stage. This charge amplification process is linear (no avalanche effects), resulting in an output charge pulse that is proportional to the incident light level. This electron multiplication process causes e propagation delay of typically $20-50$ nanoseconds, however the intrinsic spread in this delay is limited to only a few nanoseconds.

7.6.1 Photocathodes

The photo-emission process in a photocathode can be subdivided in three stages: the photon absorption in the cathode material that gives rise to the production of a photo-electron, secondly the migration of the photo-electron to the surface of the cathode and, thirdly, crossing the exit potential of the cathode material. Semiconductors are superior compared to metals as photosensitive layer regarding all three aspects:
- semiconductors provide a high photo-absorption efficiency.
- the migration length in semiconductors is larger than in metals since the electron -electron collisions that dominate in metals are practically absent in semiconductors where the interactions are mainly limited to electron−phonon scattering. This means in practice that photo-electrons can be generated over a much larger interaction depth: a few nanometers for metals as compared to a few tens nanometers for semiconductor materials.
- The vacuum exit potential for metals is relatively high, i.e. $U_M = 3-4$ eV, whereas for semiconductors this amounts to $U_s = 1.5-2$ eV.

Semiconducting photocathodes with a thickness of 20−30 nanometers are semitransparent leading to $\leq$50% absorption of the incident light. Moreover the efficiency is further reduced since the electrons excited from the valence band lose their energy in $\approx 10^{-12}$ seconds through phonon interactions, arriving at the lowest energy level in the conductance band. In case of a positive electron affinity χ, escape to vacuum is then rendered impossible. However the electron will still stay for $\approx 10^{-10}$ seconds in the conduction band before recombining. Therefore, employing a material with *negative*

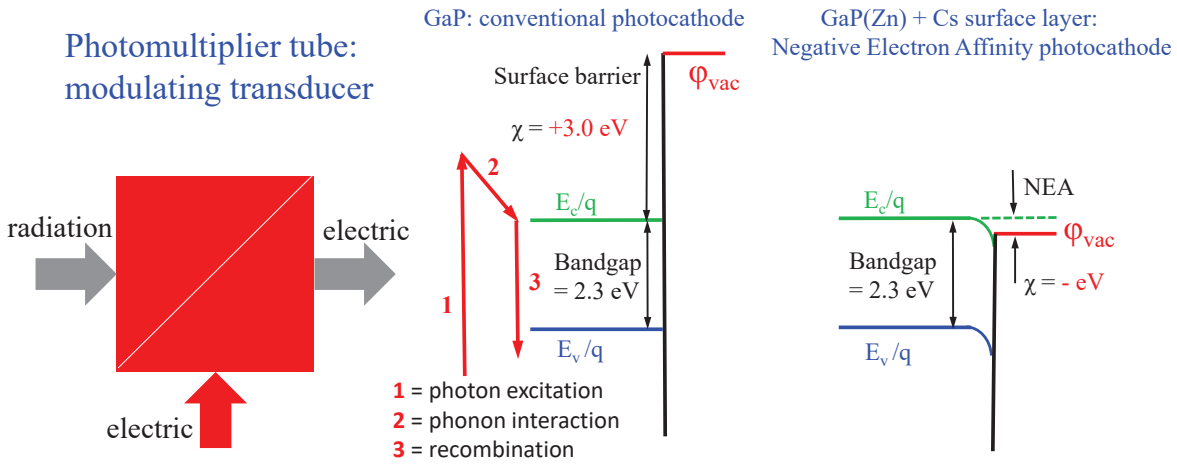


Figure 7.25: *The photomultiplier tube represents a modulating transducer (left). The right panel shows the band structure of the electron potential transition between the photocathode and the vacuum for a conventional GaP-photocathode and a GaP(Zn)+Cs-surface-layer NEA-photocathode.*

electron affinity is potentially very profitable to gain efficiency: electrons released deep in the photocathode material that are already resident on the lowest energy level of the conduction band have still 100 picoseconds rather than 1 picosecond available for escape to vacuum because $\chi < 0$. The energy band structures (i.e. electron potentials) of a conventional and a Negative Electron Affinity (NEA) photocathode are both shown in figure (7.25).
Regarding thermal electron emission, i.e. thermal noise, semiconductors are more noisy than metals: $10^6 - 10^8$ m^{-2} sec^{-1} as opposed to 10^2 m^{-2} sec^{-1} owing to the lower work function of the semiconductor. Two types of photocathodes can be distinguished:
-*opaque* photocathodes
-*semitransparent* photocathodes
Both types imply different construction: The opaque photocathodes are applied to a carrier material with a thickness that is slightly larger than the escape thickness, the photo-electrons are emitted and collected from the same side as the incident light source. The semitransparent photocathodes are applied to a transparent carrier material with a thickness that should not exceed the escape thickness. The transparent carrier is in practice the glass or quartz window of the photomultiplier. In this case the photo-electrons are collected from the side of the cathode that is opposite to the side that is illuminated by the light source.
The salient features of a photocathode comprise:

- *Efficiency*
  The detection efficiency is expressed by the quantum efiiciency QE=(number of emitted photo-electrons)/(number of incident photons). The maximum QE is 20-30% for multi-alkali material like Na_2KSb or K_2CsSb, i.e. bialkali photocathodes. A GaAs-photocathode possesses a negative electron affinity and has a high QE-value into the infrared.

- *Spectral response*
  In general the long-wavelength cut-off is determined by a low value of the photon absorption probability and the lower energy of the produced photo-electron that reduces the escape probability into the vacuum.
  At the short-wavelength end the efficiency becomes limited by the opaqueness of the window material, e.g. glass has a cut-off wavelength around 350 nanometer, which is insufficient to reach into the UV-domain. Application of quartz or fused silica broadens the transmission to wavelengths as low as 160 nanometer. The spectral response of a photocathode is normally specified by an S-number code, examples are S11 (visible range) and S20 (near-IR response).

- *Uniformity*
  The uniformity in the thickness of the photosensitive layer across the cathode area is an important parameter for the achievable energy resolution in radiation detectors. Variation in thickness gives rise to proportional variations in sensitivity of different positions on the photocathode area. This leads to different charge levels for the same incident radiation flux, depending on photocathode position. Cathode non-uniformity is especially a problem with phototubes with a large diameter, where also the collection efficiency of photo-electrons onto the first dynode for electron multiplication plays a role. The latter is dependent on the quality of the electron-optics between the photocathode and the series of dynode stages.

7.6.2 Electron multiplier

The electron multiplier comprises a number of electrodes (dynodes) with mutually increasing potential that energizes the incoming photo-electrons that much by acceleration that each electron striking the dynode material gives rise to the production of a number of secondary electrons which results in charge amplification. A photo-electron leaving the cathode has a kinetic energy of ≈ 1 eV, hence the acceleration potential determines the the impact energy on the dynode surface. The secondary emission yield per incident electron is a sensitive function of the impact energy: a low kinetic energy only produces secondaries in the surface layer but with a large probabilty of escape, a high kinetic energy excites more secondaries but deeper in the dynode material and have as a consequence a lower escape probability. This necessitates selection of an optimized impact energy for maximum secondary electron yield. The multiplication factor (or secondary emission factor) is defined as $\delta =$ (number of emitted secondary electrons)/(number of

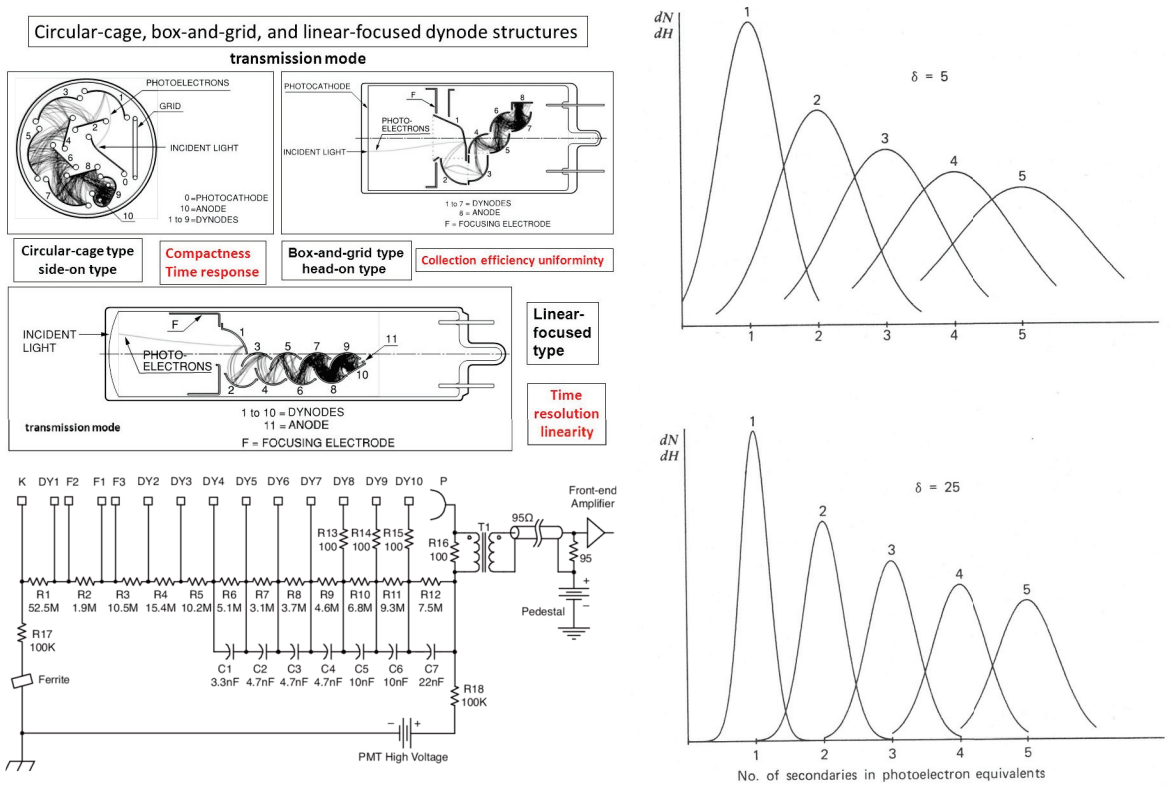


Figure 7.26: *Three different dynode configurations (types) in a photomultiplier tube. A voltage divider for distribution of the High Voltage (HV) over the dynode stages is also shown. As can be seen the potential difference between the photocathode and the first dynode is much larger than for the other dynode stages. A NEA-dynode can raise the secondary emission factor δ from $\approx$ 5 to $\geq$25. In the panel at the right hand side the improvement in resolution with a NEA first dynode is displayed expressed in 1, 2, 3, 4 and 5 photo-electron equivalents respectively.*

incident primary electrons).
A very important development has been the application of materials with negative electron affinity (NEA) as dynode surface layer. The most successful of these materials is Gallium-Phosphide (GaP) with a high impurity concentration of Zn (p^+-semiconductor, 10^{25} m^{-3}). At the surface of this semiconductor a very thin, almost mono-atomic, layer of an electron positive material is applied, in this case Cesium (Cs). The Zn-acceptor impurities become filled with these electrons resulting in an ionized layer of positively charged Cs-atoms at the surface that are bound by electrostatic forces. This causes downward bending of the conduction and valence band at the material surface owing to the negatively charged Zn-ions that constitute a dipole layer with the positively charged layer of Cs-ions. As a result of the bending, the vacuum electron potential ends up below the lower potential boundary level of the conduction band, see again figure (7.25). In this way the photon-excited electrons that energetically reside during $\approx$ 100 picoseconds at the lower boundary of the conduction band have still a substantial probability for escape. The NEA-characteristic has the following advantages:
- The secondary emission yield keeps increasing with the primary electron impact en-

ergy, since the deeper layers can also contribute. A yield $\delta =$ 50-60 at a potential of 1000 V is feasible for GaP(Cs).
-Ultra-fast timing becomes more accurate since the energy distribution of the secondaries is much narrower than in the case of conventional dynode materials. The reason is that practically all produced secondary electrons were initialized from the same potential level, i.e. the bottom potential of the conduction band. Moreover, to get to the same level of charge amplification as with conventional dynode surface layers, less stages are required which reduces the spread in the propagation delay.
- High values of δ give a lower statistical spread in the ultimate charge signal, which potentially allows for resolving the single photo-electron distribution (capability for single photon counting).

The charge amplification and the associated statistical fluctuations can be assessed in the following way. The total charge amplification can be expressed as:

$$G = \alpha \cdot \delta^N \quad \alpha = \text{number of photo-electrons, } N = \text{number of dynodes} \tag{7.131}$$

For G=10^7 and δ =5, 10 stages will be necessary, for NEA dynodes with a δ =55 only 4 stages do suffice. In case $\delta \propto (\Delta V)_{dyn} \propto V$, a 10 stage phototube would have $G \propto V^{10}$. However δ only increases with an exponent of V smaller than one, therefore in reality we have $G \propto V^{6-9}$.
The secondary emission factor δ is a random variable, in the simplest model it can be assumed that it exhibits a Poissonian probability distribution with mean δ and standard deviation $\sigma = \sqrt{\delta}$. The relative variance $(\sigma/\delta)^2 = 1/\delta$ and considering N dynode stages, the total relative variance can be expressed as:

$$\sum_{i=1}^{N} \delta^{-i} \approx \frac{1}{\delta - 1} \tag{7.132}$$

If $\delta \gg 1$, the relative variance is almost entirely governed by the yield of the first dynode. Figure (7.26) shows the large improvement that can be achieved by depositing NEA-material on the surface of the first dynode. Discrimination against thermal noise pulses (single photo-electrons) and the ability to perform single photon counting are evident.
In reality the probability distribution functions turn out to be broader than a Poissonian distribution, even exponential distributions with no peak are encountered. Assessment of the multiplication statistics with more realistic distribution functions often employ so-called Polya distributions, these allow variation between Poisson and exponential distributions by the choice of a single parameter.

7.6.3 Noise Equivalent Power (NEP) and limiting sensitivity

Thermionic emission from the photo-cathode is the predominant source of dark current in photomultiplier (PM) tubes, the magnitude of which is given by Richardson's formula that was presented in Chapter 4 in equation (4.4). Thermionic emission by the dynodes can also contribute to the dark current, but the photo-cathode usually dominates

over the dynode contribution since this photo-current is subject to the largest charge amplification. Richardson's equation shows the thermionic emission to be proportional to the square of the temperature, consequently cooling of the PM-tubes will effectively reduce the dark current and will enlarge the linear dynamic range of the device at the small-signal end.
A photomultiplier tube is essentially a vacuum diode and hence shot noise is produced by the variation in the flow of photo-electrons. The shot noise produced at the photo-cathode, see equation (5.62), amounts to:

$$\overline{\Delta I_C^2} = 2q\,\overline{I}\Delta\nu_c,\ \Delta\nu_c = \text{noise power bandwidth},\ = 2q\,(\overline{I}_{ph} + \overline{I}_{dark})\Delta\nu_c \tag{7.133}$$

Assuming that the PM-tube performance is shot noise limited by the dark current, we can write the following equation for derivation of the NEP:

$$q\eta\phi_{ph} = q\eta\phi_E\frac{\lambda}{hc} = (2q\overline{I}_{dark}\Delta\nu_c)^{\frac{1}{2}} \ \Rightarrow\ \text{NEP} = \phi_E = \frac{hc}{q\eta\lambda}\left(2q\overline{I}_{dark}\Delta\nu_c\right)^{\frac{1}{2}} \tag{7.134}$$

At the PM-tube anode the total noise includes the effect of the charge amplification and the thermal noise contribution from the resistive voltage divider that sets the dynode potentials. Hence we get:

$$\sqrt{\overline{\Delta I_A^2}} = \left[2qG^2\,(\overline{I}_{ph} + \overline{I}_{dark})\Delta\nu_c + \frac{4kT\Delta\nu_c}{R_{div}}\right]^{\frac{1}{2}} \tag{7.135}$$

From equation (7.135) it becomes evident that for sufficiently large values of the charge gain G and the divider resistance R_{div}, the PM-tube performance will be shot noise limited. Of course the value of R_{div} should not be so high that the anode current would start to influence the charge gain, but the value does have to be chosen as high as possible to reduce the thermal resistor noise as much as possible. In practice it is mostly possible to reach shot noise dominated performance and this being the case, the signal to noise ratio (SNR) at the anode is not affected by the charge gain factor G if it is kept constant:

$$SNR_A = \frac{G\overline{I}}{\sqrt{2qG^2\,(\overline{I}_{ph} + \overline{I}_{dark})\Delta\nu_c}} = \frac{\overline{I}}{\sqrt{2q\,(\overline{I}_{ph} + \overline{I}_{dark})\Delta\nu_c}} \tag{7.136}$$

The above equation shows that in the shot noise limited case, the SNR at the photocathode and at the PM-anode are equal, in the absence of a dark current the SNR becomes governed by signal-photon-limited noise, however degraded by the quantum efficiency factor η. Given a photon count integration period ΔT, the incident SNR_{in} equals $\sqrt{\overline{\phi}_{ph}\Delta T}$, the resulting SNR for the outgoing photo-current follows then from:

$$SNR_{out} = \sqrt{\eta}SNR_{in} \tag{7.137}$$

7.7 Bolometer: thermal sensing

7.7.1 Operation principle

The operation principle for a bolometer is that a change in temperature, that is produced by the absorption of incident radiant power, causes a change in the electrical

resistance of the material used to fabricate the bolometer itself or of 'thermometer' material that has a strong thermal link with the bolometer. So the interaction processes with the incoming radiation field do not release free charge carriers but result in the production of heat, i.e. lattice vibrations, which in turn produce the resistance change in the material. The resistance change of a material can be expressed by a linear equation, introducing a temperature coefficient of resistance α:

$$R = R_0(1 + \alpha\Delta T), \quad \text{with} \quad \alpha = \frac{1}{R}\frac{dR}{dT}, \quad [\text{Kelvin}]^{-1} \tag{7.138}$$

where R_0 represents the resistance at some reference temperature.
Bolometers can be divided into five different types

- *Metal bolometers*
  Made of metal, these bolometers need to be small so that their heat capacity C is low enough to allow reasonable sensitivity. They are mostly fabricated as thin strips ($10-50$ nanometer) by vacuum evaporation or sputtering. The temperature coefficient α is of the order 0.5%/K, hence it is positive and implies resistance increase with temperature.

- *Thermistors*
  The temperature sensitive material of a thermistor typically comprises sintered wafers of metal oxides like Mn-, Ni-, and Co-oxides that are mounted on an electrically isolating but thermally conductive substrate like sapphire that acts as a heat sink. When the temperature of the oxide film increases by radiation absorption, the concentration of thermally excited free charge carriers in the oxide increases, hence the resistance decreases. Consequently, the coefficient of resistance α is negative in this case and typically amounts to 5%/K, but does vary as $1/\mathrm{T}^2$. For this reason, a *'matched pair'* is used as a single unit in practice: one thermistor of the pair is shielded from the incoming radiation and fitted in an electrical bridge such that it acts as a matched load resistor providing maximum signal transfer over ambient temperatures.

- *Semiconductor bolometers*
  Lightly-doped Germanium bolometers belong to the most used semiconductor bolometers. In particular, when operated at cryogenic temperatures, performance near the theoretical limit can be achieved over a wavelength range ranging from a few to a approximately 100 micron, i.e. in the mid- and far-infrared range. In general, semiconductors when lightly doped with suitable impurities constitute excellent bolometer material. Ga-doped Germanium, cut from single crystal Germanium, can be easily lapped and etched to a desired thickness, the thermal conductivity of Germanium is sufficiently high that a uniform temperature can be secured between the center and the edges of the bolometer element. An empirical relation for the resistance versus temperature of Ge-bolometric material has been established at cryogenic temperatures between 1 and 5 K: given a resistance R_0

at a temperature T_0, the following relations hold:

$$R(T) = R_0 \left(\frac{T_0}{T}\right)^{\gamma}, \quad \text{with } \gamma \approx 4 \Rightarrow \quad \alpha(T) = \frac{1}{R}\frac{dR}{dT} = -\frac{\gamma}{T} \tag{7.139}$$

Equation (7.139) shows that the negative temperature coefficient α is inversely proportional to the temperature, which is advantageous for the responsivity of such a bolometer towards lower temperatures. Moreover the specific heat C/m, m representing the bolometer thermal mass, also reduces when the temperature falls. In the next section we shall describe two bolometer bias configurations, i.e. fixed current bias and fixed voltage bias respectively, and from these configurations we shall derive expressions for the associated bolometer responsivities.

- *Composite bolometers*
  The concept of a composite bolometer was inspired by the wish to lower the heat

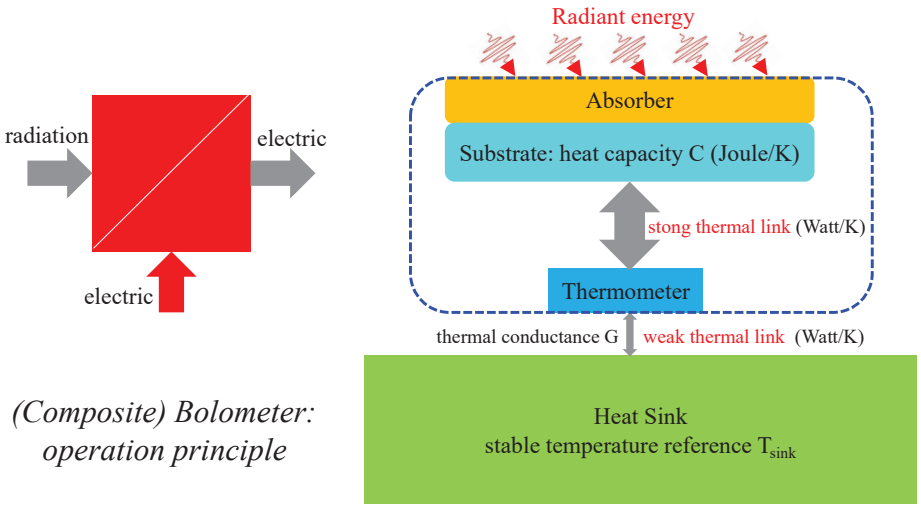


Figure 7.27: *A (composite) bolometer is a modulation transducer that performs thermal sensing. It constitutes an absorption layer in which the incident radion power is converted into heat that raises the temperature of the absorber/substrate combination with an amount ΔT depending on its heat capacity C [Joule/Kelvin]. A (much smaller) thermometer element that is thermally strongly linked to the absorber/substrate layer converts this temperature rise into a current (voltage) signal. The heat balance is restored through a weak thermal conduction link G [Watt/Kelvin] with a constant temperature reference (heat sink). The blue dashed line indicates integration of the constituting elements on a single chip.*

capacity of the device to reduce the time constant $\xi = C/G$ and enhance the temporal frequency response. For instance the heat capacity of a sapphire layer, a perfect electrical insulator with good thermal conductance, possesses about 1/50th of the heat capacity of a Germanium layer of similar dimension which implies that a much larger sensor area can be realized without loss of speed.
A composite bolometer consists of three parts, a sketch is given in figure (7.27).

The absorber (1) can be made of a thin film of an appropriate material with very high emissivity in the desired waveband, e.g. in the far-infrared above a hundred micron Black Bismuth and Nichrome absorbers are often used. The thin film absorber is deposited on a substrate with low heat capacity and good thermal conductivity, e.g. sapphire. Subsequently a very much smaller temperature sensor (e.g. Germanium), *the thermometer* (2), is bonded with epoxy or varnish of excellent thermal conductivity to the film/substrate combination. The small, low heat capacity, thermometer is embedded in an electrical bias circuit that converts a temperature change into an electrical signal, it also provides a weak thermal link to a *heat sink* (3) that provides a stable temperature reference. In this way a relatively large and effective absorbing element is combined with a low-heat-capacity thermometer read-out of small dimension, while still maintaining an acceptable frequency bandwidth.

- Superconducting bolometer
  The composite bolometer specifically finds its application in low-noise, low-

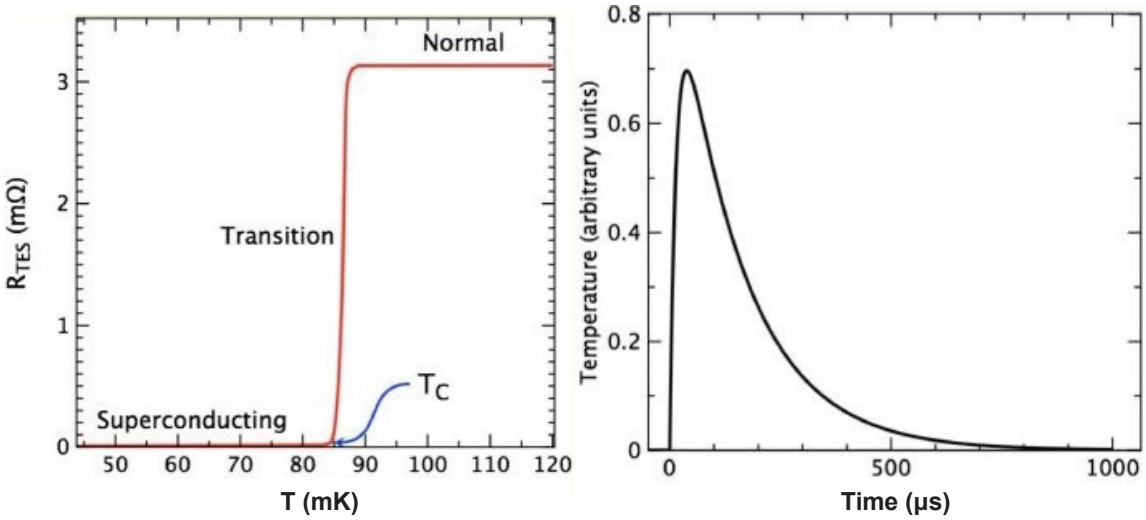


Figure 7.28: *Left panel: thermometer edge-like resistance transition at the critical temperature T_c from the superconductive to the normal conductive state. Right panel: typical heat pulse in response to a full energy deposit by a single X-ray photon, demonstrating the bolometer capability for single photon counting with high energy resolution. The decay time of the temperature pulse is governed by the ratio between the heat capacity C of the thermometer and the heat conduction link G from the thermometer to the heat sink: $\xi = C/G$.*

temperature (cryogenic) infrared and far-infrared sensing. However a new type of composite bolometer for this purpose has been introduced over the last decade that operates on the conductivity transition of a suitable superconducting material that steeply changes its resistance over orders of magnitude going from super- to normal conductivity, see figure (7.28) left panel. If the temperature of the bolometer is held near the midpoint of this transition, energy deposited by

an incoming radiation field can produce sufficient heat to cause large resistance changes. This application requires a very stringent temperature control to stably maintain the midpoint setting on the resistance transit curve. This bolometer type is also used for single photon detection in the X-ray waveband, each absorbed incident photon produces a heat pulse with a typical characteristic shape shown in figure (7.28) right panel. The operation of suitable materials for these so-called Transition Edge Sensor (TES) microbolometers requires cryogenic temperatures in the tens of milli-Kelvin range, see figure. This detection technique is now successfully developed for space applications to accommodate high resolution spectral measurements of cosmic X-ray sources with an energy-dispersive rather than a wavelength-dispersive method.

7.7.2 Bolometer bias configurations and responsivity

Bias circuits that are used to interface the bolometer with a temperature sensing system are similar to those employed for a photoconductor or make use of a Wheatstone bridge. The latter configuration is mainly applied for thermistor read-out: two thermistors are electrically balanced in a bridge when no radiation is present, subsequently only one thermistor is illuminated by a radiant flux causing a differential current to flow following the unbalance of the bridge caused by the resistance change of the irradiated device.
Assessment of the temperature change ΔT requires now a more complicated expression for the heat equation that we introduced in (5.91) for computing the phonon noise. A bias current (I_b) flowing through the bolometer (or thermometer), the connecting wires and the resistive load generates *Joule heating* $I_{bias}^2 R_b$ (R_b = bolometer resistance) in the read-out circuitry. We will treat here two typical cases: the bolometer and a load resistance are either fed by a fixed current source or by a fixed voltage source. Both cases are displayed in figure (7.29). The fixed current source configuration is often used with cryogenic semiconductor (e.g. Germanium) bolometers where the load resistance $R_L \gg R_b$. The voltage bias configuration is often used in the case of balanced thermistors when $R_L = R_b$

- *Fixed-current-source bias*
  Let's first examine the the fixed current source bias configuration shown in figure (7.29), left panel. Assigning to the bolometer a temperature T_b and to a heat sink a temperature T_{sink} in the absence of a radiant flux we have the following equilibrium condition in the heat equation:

$$G(T_b - T_{sink}) = I_B^2 R_b \tag{7.140}$$

  If a background radiation field is present, generating an energy flux ϕ_{bkd} incident on the bolometer, this can be taken into account by defining an effective sink temperature T_{sink}^* according to $G(T_{sink}^* - T_{sink}) = \epsilon\phi_{bkd}$ (ϵ = detection efficiency), potentially introducing some degradation in performance. However, let's assume for now that $\epsilon\phi_{bkd} \ll I_B^2 R_b$, we put $T_{sink}^* = T_{sink}$ and, writing $P = I_B^2 R_b$, we arrive at the following equation for the differential heating ΔT owing to an incoming

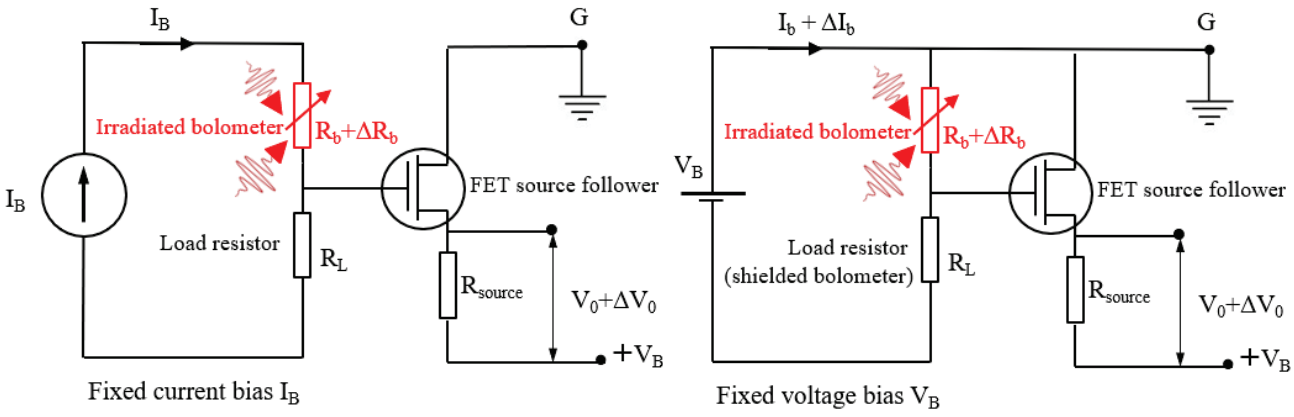


Figure 7.29: *Fixed current bias (left side) and fixed voltage bias (right side) circuits for a bolometer with a FET source follower read-out. In case of a thermistor it is common to use a shielded thermistor of the same characteristics as load impedance ($R_L = R_b$), which secures maximum responsivity.*

radiant signal flux:

$$C\frac{d\Delta T}{dt} + G\Delta T = \frac{dP}{dT}\Delta T + \epsilon\phi_{sig}, \tag{7.141}$$

And:

$$\frac{dP}{dT} = \frac{d(I_B^2 R_b)}{dT} = I_B^2\frac{dR_b}{dT} = \alpha R_b I_B^2 \tag{7.142}$$

We can now define an effective heat conduction coefficient:

$$G_e \;=\; G - \alpha R_b I_B^2 = G[1 - \alpha(T_b - T_{sink})] \tag{7.143}$$

and express the heat equation as:

$$C\frac{d\Delta T}{dt} + G_e\Delta T = \epsilon\phi_{sig} = \epsilon\phi_0 e^{j\omega t}, \tag{7.144}$$

where $\phi_0 e^{j\omega t}$ represents a periodic radiant signal. This differential equation (7.144) can be easily solved with the Lagrange method (variation of constants), its solution comprises two terms:

$$\Delta T \;=\; A\, e^{-(t/\tau_{e_{th}})} + \frac{\epsilon\phi_0 e^{j\omega t}}{G_e(1 + j\omega\tau_{e_{th}})}, \tag{7.145}$$

with $\tau_{e_{th}} = C/G_e = C/(G - \alpha R_b I_B^2) = \tau_{th}/[1 - \alpha(T_b - T_{sink})]$ the *effective* thermal time constant of the bolometer and A a numerical constant.
If G_e is positive the first term in solution (7.145) will exponentially decay to zero, however if G_e is negative the first term will exponentially increase and burn-out of the bolometer might occur. 'Burn-out' can occur when the heat dissipated into the bolometer exceeds the heat conducted away by the electrical leads. To prevent this from happening the following condition needs to be satisfied:

$$G - \alpha R_b I_B^2 \;>\; 0 \tag{7.146}$$

This condition is in any case satisfied for thermistors and (cryogenic) semiconductor bolometers, since the temperature coefficient of resistance α is negative for those cases.
From the second term in equation (7.145) we can write the persistent (absolute) magnitude of the temperature increase as:

$$|\Delta T| = \frac{\epsilon\phi_0}{G_e(1+\omega^2\tau_{e_{th}}^2)^{1/2}} \tag{7.147}$$

For the resistance change of the bolometer and the associated voltage change we now arrive at:

$$\begin{aligned} \Delta R_b &= \alpha R_b|\Delta T| \quad \Rightarrow \quad \Delta V_b = I_B\Delta R_b = \frac{\alpha\epsilon R_b\phi_0}{G_e(1+\omega^2\tau_{e_{th}}^2)^{1/2}}I_B, \\ \text{and } \Delta V_0 &= \frac{\alpha\epsilon\phi_0}{G_e(1+\omega^2\tau_{e_{th}}^2)^{1/2}}\frac{R_bR_L}{(R_b+R_L)}I_B \end{aligned} \tag{7.148}$$

For the voltage responsivity $R_V = \Delta V_0/\phi_0$ of the bolometer we get:

$$R_V = \frac{\alpha\epsilon}{G_e(1+\omega^2\tau_{e_{th}}^2)^{1/2}}\frac{R_bR_L}{(R_b+R_L)}I_B = \frac{\alpha\epsilon(R_b//R_L)}{G_e(1+\omega^2\tau_{e_{th}}^2)^{1/2}}I_B \tag{7.149}$$

If $R_L \gg R_b$ the voltage responsivity reduces to:

$$R_V = \frac{\alpha\epsilon}{G_e(1+\omega^2\tau_{e_{th}}^2)^{1/2}}V_b \quad \text{with } V_b \text{ the voltage over the bolometer,} \tag{7.150}$$

which simplifies, in case $\epsilon \approx 1$ and the roll-off term $\omega^2\tau_{e_{th}}^2$ can be neglected, to:

$$R_V = \frac{\alpha}{G_e}V_b = \frac{\alpha V_b}{G - \alpha I_B^2 R_b} \tag{7.151}$$

Using condition (7.140) for heat equilibrium, we can also obtain the electrical load equations as a function of the bolometer temperature T_b and the associated resistance $R_b(T_b)$:

$$V_b(T_b) = \sqrt{G\,R_b(T_b)(T_b - T_{sink})} \tag{7.152}$$

$$I_b(T_b) = \sqrt{\frac{G\,(T_b - T_{sink})}{R_b(T_b)}} \tag{7.153}$$

By substituting V_b in equation (7.151) we can express the responsivity R_V as a function of the sensor temperature (T_b) and the bath temperature (T_{sink}):

$$R_V(T_b, T_{sink}) = \frac{\alpha\sqrt{R_b(T_b)G(T_b - T_{sink})}}{G[1-\alpha(T_b - T_{sink})]} = \frac{\alpha\sqrt{T_b - T_{sink}}}{1-\alpha(T_b - T_{sink})}\sqrt{\frac{R_b(T_b)}{G}} \tag{7.154}$$

We shall employ equation (7.154) later on when evaluating the noise equivalent power (NEP) due to phonon and to Nyquist noise in cryogenic Ge-bolometers.

By taking the Fourier transform of equation (7.150), we arrive at the responsivity in the time domain:

$$R_V = \frac{\alpha\epsilon}{G_e}(1 - e^{-t/\tau_{e_{th}}})V_b \tag{7.155}$$

It is clear from this expression that if the effective thermal conductance is increased to gain frequency response, the responsivity goes down, i.e. the values for responsivity and frequency response of a bolometer are anti-correlated.

- *Fixed-voltage-source bias*
  Figure (7.29) right panel shows the bolometer embedded in a bias circuit with a fixed supply voltage V_B. This configuration is commonly utilized for thermistor read-out. Taking a bolometer temperature T_b and a heat sink temperature T_{sink} in the absence of a radiant flux, equilibrium for heat exchange dictates:

$$G(T_b - T_{sink}) = I_b^2 R_b = V_B^2 \left[\frac{R_b}{(R_b + R_L)^2}\right] \tag{7.156}$$

  Furthermore we can write for the electrical power dissipation:

$$\frac{dP}{dT} = \frac{d(I_b^2 R_b)}{dT} = V_B^2 \frac{d}{dT}\left[\frac{R_b}{(R_L + R_b)^2}\right] = \alpha\, G\, (T_b - T_{sink})\frac{R_L - R_b}{R_L + R_b} \tag{7.157}$$

  The effective heat conduction coefficient can now be expressed as:

$$G_e \;=\; G - \alpha\, G\, (T_b - T_{sink})\frac{(R_L - R_b)}{(R_L + R_b)}, \tag{7.158}$$

  simplifying, as before, the heat equation to:

$$C\frac{d\Delta T}{dt} + G_e \Delta T = \epsilon\phi_{sig} = \epsilon\phi_0 e^{j\omega t}, \tag{7.159}$$

  Solving this differential equation yields the resistance change of the bolometer as:

$$\begin{aligned}
\Delta R_b &= \alpha R_b \Delta T, \;\text{ from the bias circuit in figure (7.29) we have:} \\
V_0 &= -\frac{R_L}{R_L + R_b} V_B \;\Rightarrow\; \Delta V_0 = \frac{R_L \Delta R_b}{(R_L + R_b)^2} V_B \;\Rightarrow \\
\Delta V_0 &= \frac{\alpha\epsilon\phi_0}{G_e(1 + \omega^2\tau_{e_{th}}^2)^{1/2}} \frac{R_L R_b}{(R_L + R_b)^2} V_B
\end{aligned} \tag{7.160}$$

  Hence, for the voltage responsivity $R_V = \Delta V_0/\phi_0$ of the bolometer we get:

$$R_V = \frac{\alpha\epsilon}{G_e(1 + \omega^2\tau_{e_{th}}^2)^{1/2}} \frac{R_L R_b}{(R_L + R_b)^2} V_B \tag{7.161}$$

  Maximum responsivity can be achieved for the special case that $R_b = R_L$. This is the case for a thermistor pair: R_L represents a shielded thermistor that has the same physical characteristics as the illuminated one:

$$R_V = \frac{\alpha\epsilon}{4G_e(1 + \omega^2\tau_{e_{th}}^2)^{1/2}} V_B \tag{7.162}$$

  By taking the Fourier transform of equation (7.162), we can again write the responsivity in the time domain:

$$R_V = \frac{\alpha\epsilon}{4G_e}(1 - e^{-t/\tau_{e_{th}}})V_B \tag{7.163}$$

7.7.3 Limiting sensitivity

To assess the limiting sensitivity of a bolometer we need to accumulate all contributing noise components, i.e. the phonon noise, the thermal (Nyquist) noise from the resistive elements, the $1/f$-noise and the equivalent current and voltage noise sources that can be attributed to the preamplifier. This can be expressed in the noise equivalent power (NEP) by utilizing the voltage responsivity.
The ultimate limiting factor in performance of any thermal detection device is the phonon noise. The spectral NEP$(\lambda, \Delta\nu_N)$ at a particular wavelength λ integrated over a noise bandwidth $\Delta\nu_N$ for a bolometer operated with a fixed current source equals:

$$NEP(\lambda, \Delta\nu_N) = \frac{\sqrt{\overline{\Delta V^2}}}{R_V(\lambda, \Delta\nu_N)} \tag{7.164}$$

We can write:

$$\sqrt{\overline{\Delta V_N^2}} = I_B\sqrt{\overline{\Delta R_b^2}} = \alpha I_B R_b\sqrt{\overline{\Delta T^2}} = \alpha V_b\sqrt{\overline{\Delta T^2}}, \tag{7.165}$$

in which we can substitute the expression (5.93) for $\overline{\Delta T^2}$ that we derived in chapter 5 when we discussed phonon noise:

$$\sqrt{\overline{\Delta V_N^2}} = \alpha V_b\sqrt{\frac{4kT^2}{G_e}\Delta\nu_N} \tag{7.166}$$

Substitution of the above expressions for $R_V(\lambda, \Delta\nu_N)$ and $\sqrt{\overline{\Delta V_N^2}}$ in (7.164) yields:

$$NEP(\lambda, \Delta\nu_N) = \sqrt{4kT^2G_e\Delta\nu_N}\cdot\frac{(1+\omega^2\tau_{e_{th}}^2)^{1/2}}{\epsilon(\lambda)}\frac{(R_L+R_b)}{R_L} \quad \text{[Watt]} \tag{7.167}$$

In many practical cases, employing a fixed current source, we have $R_L \gg R_b$. In the case of $\epsilon(\lambda) \approx 1$ and a frequency domain where the roll-off term $\omega^2\tau_{e_{th}}^2$ can be disregarded, (7.167) reduces to (5.93):

$$NEP(\Delta\nu_N) = \sqrt{4kT^2G_e\Delta\nu_N} = \sqrt{S_{P_{th}}(\nu)\Delta\nu_N} \quad \text{[Watt]} \tag{7.168}$$

Accounting for the full noise bandwidth of the bolometer we can substitute $\Delta\nu_N = G_e/4C = 1/4\tau_{e_{th}}$, and still have in good approximation:

$$NEP(\Delta\nu_{e_{th}}) = \sqrt{\frac{kT^2G_e}{\tau_{e_{th}}}} \quad \text{[Watt]} \tag{7.169}$$

A similar equation can be derived for the case that the bolometer is fed by a fixed voltage source V_B. In this case we have the following expressions:

$$\begin{aligned}\sqrt{\overline{\Delta V_N^2}} &= \frac{\alpha R_b}{R_b+R_L}V_B\sqrt{\overline{\Delta T^2}} = \frac{\alpha R_b}{R_b+R_L}V_B\sqrt{\frac{4kT^2}{G_e}\Delta\nu_N}\\ R_V(\lambda, \Delta\nu_N) &= \frac{\alpha\epsilon(\lambda)}{G_e(1+\omega^2\tau_{e_{th}}^2)^{1/2}}\frac{R_LR_b}{(R_L+R_b)^2}V_B\end{aligned} \tag{7.170}$$

Substituting again in (7.164) yields:

$$NEP(\lambda, \Delta\nu_N) = \sqrt{4kT^2G_e\Delta\nu_N} \cdot \frac{(1+\omega^2\tau_{e_{th}}^2)^{1/2}}{\epsilon(\lambda)} \frac{(R_L+R_b)}{R_L} \tag{7.171}$$

If a thermistor pair is employed we have $R_b = R_L$, taking again $\epsilon(\lambda) \approx 1$ and considering a frequency range where the roll-off term $\omega^2\tau_{e_{th}}^2$ can be neglected, (7.171) reduces to:

$$NEP(\Delta\nu_N) = \sqrt{16kT^2G_e\Delta\nu_N} = 4\sqrt{kT^2G_e\Delta\nu_N} \quad \text{[Watt]} \tag{7.172}$$

Taking the full intrinsic bandwidth for $\Delta\nu_N$, we obtain in good approximation:

$$NEP(\Delta\nu_{e_{th}}) = 2\sqrt{\frac{kT^2G_e}{\tau_{e_{th}}}} \quad \text{[Watt]} \tag{7.173}$$

The difference of a factor two with (7.169) can be attributed to different assumptions regarding the load configurations, i.e. $R_L \gg R_b$ versus $R_L = R_b$.
From this analysis it is clear that the intrinsic NEP of a bolometer solely depends on the sink temperature and the value of the effective thermal conductance G_e. Since the thermal conductance is determined by the bolometer leads , it is independent of the dimensions of the sensor element. So, unlike most radiation sensors, the NEP does not vary as the square root of the sensor area. This means that the specific detectivity D^* cannot be used as a valid parameter for inter-comparison with other sensors.

7.7.4 Cryogenically cooled Germanium bolometer

To properly assign a correct temperature to the NEP given the difference in temperature between the electrothermally heated bolometer (T_b) and the bath temperature (T_{sink}) we shall elaborate here, *as an example*, on the NEP of a cryogenic Ge-bolometer. An empirical relationship between resistance and temperature for this sensor was established in the 1.1-4.2 Kelvin range by Frank Low and collaborators in 1961. We already alluded to this relationship in the first section on operational principle and repeat it for convenience here:

$$R(T) = R(T_0)\left(\frac{T_0}{T}\right)^{\gamma}, \quad \alpha = \frac{1}{R}\left(\frac{dR}{dT}\right) = -\left(\frac{\gamma}{T}\right), \quad \gamma \approx 4 \tag{7.174}$$

Next, we can substitute these relations in expression (7.154) by taking $T = T_b$ and $T_0 = T_{sink}$ and by defining the ratio $\zeta = T_b/T_{sink}$ and writing eventually the responsivity as a function of ζ:

$$\begin{aligned} R_V(T_b, T_{sink}) &= -\frac{(\gamma/T_b)\sqrt{T_b - T_{sink}}}{1+(\gamma/T_b)(T_b-T_{sink})}\sqrt{\frac{R_b(T_b)}{G}} = \\ &= -\frac{\sqrt{(T_{sink}/T_b)^{\gamma}(\gamma/T_b)^2(T_b-T_{sink})T_{sink}}}{1+(\gamma/T_b)(T_b-T_{sink})}\sqrt{\frac{R_b(T_{sink})}{GT_{sink}}} \quad \Rightarrow \\ R_V(\zeta) &= -\sqrt{\frac{\gamma^2(\zeta-1)}{[(\gamma+1)\zeta-\gamma]^2\zeta^{\gamma}}}\sqrt{\frac{R_b(T_{sink})}{G\,T_{sink}}} \end{aligned} \tag{7.175}$$

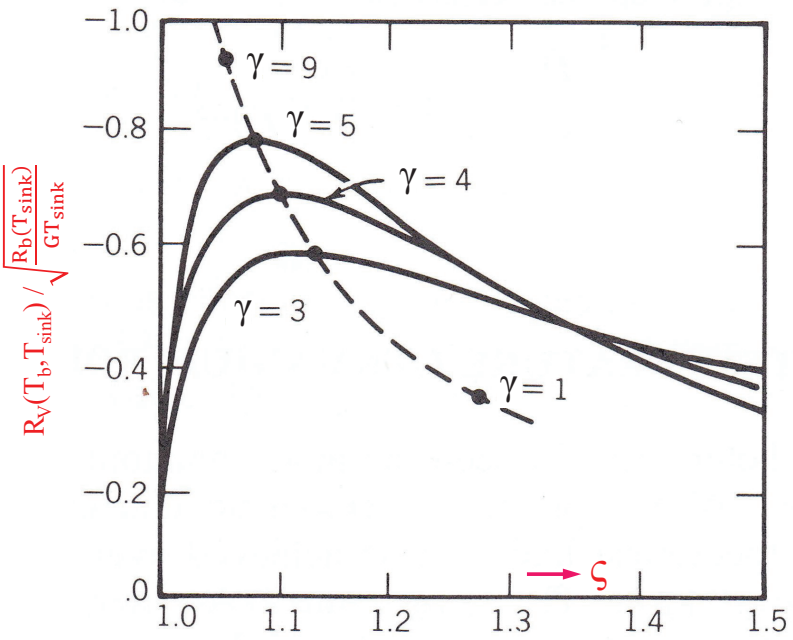


Figure 7.30: *Plot of equation (7.175) for a few distinct values of an empirically determined constant γ. The dashed line shows the position of the responsivity maxima for other value of γ (after Low 1963).*

This expression shows that given a certain bath temperature T_{sink}, the responsivity exhibits a maximum value that is only dependent on the value of the constant γ. Figure (7.30) displays equation (7.175) normalized to $\sqrt{R_b(T_{sink})/(G\,T_{sink})}$ for a few values of γ. If we select $\gamma = 4$ from the empirical relation mentioned above, the maximum value of the responsivity $R_V(\zeta)$ occurs around $\zeta = 1.1$, i.e. at $T_b = 1.1\,T_{sink}$ with a normalized value of ≈ -0.7, so we obtain a responsivity (at low frequencies) of:

$$R_V(Ge_{cryo}) = -0.7\sqrt{\frac{R_b(T_{sink})}{G\,T_{sink}}} \tag{7.176}$$

The $\mathrm{NEP}_{Ny}(Ge_{cryo})$ due to thermal (Nyquist) noise of the bolometer resistance $R_b(T_b)$ at temperature T_b can now be calculated from:

$$\begin{aligned} NEP_{Ny}(Ge_{cryo}) &= \frac{\sqrt{4kT_bR_b(T_b)\Delta\nu_N}}{R_V(Ge_{cryo})} = \\ &= \frac{\sqrt{4k\cdot\left(\frac{T_b}{T_{sink}}\right)\cdot T_{sink}\cdot R_b(T_{sink})\cdot\left(\frac{T_{sink}}{T_b}\right)^{\gamma}\cdot\Delta\nu_N}}{0.7\sqrt{R_b(T_{sink})/(G\,T_{sink})}} \approx \\ &\approx \frac{\sqrt{3kT_{sink}^2G\Delta\nu_N}}{0.7} = 2.45\sqrt{kT_{sink}^2G\Delta\nu_N} \end{aligned} \tag{7.177}$$

and the phonon noise equivalent power:

$$NEP_{ph}(Ge_{cryo}) = \left(\frac{-1}{-0.7}\right)\sqrt{\frac{\gamma^2 \cdot 4kT_b^2 \cdot GR_b(T_b)(T_b - T_{sink}) \cdot GT_{sink} \cdot \Delta\nu_N}{T_b^2 \cdot G[1 + (\gamma/T_b)(T_b - T_{sink})] \cdot R_b(T_{sink})}} = \\ = \left(\frac{1}{0.7}\right)\sqrt{\frac{\gamma^2(\zeta - 1)}{[(\gamma + 1)\zeta - \gamma]\zeta^{\gamma-1}}}\sqrt{4kT_{sink}^2 G\Delta\nu_N} \quad (7.178)$$

substituting $\alpha = 4$ and $\zeta = (T_b/T_{sink}) = 1.1$ this results in:

$$NEP_{ph}(Ge_{cryo}) \approx 2.56\sqrt{kT_{sink}^2 G\Delta\nu_N} \quad (7.179)$$

Comparison of the NEP values due to Nyquist noise (7.177) and phonon noise (7.179) shows that they are similar in magnitude for a T_b/T_{sink} ratio that provides maximum responsivity at the relevant value of the constant γ.

7.8 Pyro-electric sensor element: thermal sensing

7.8.1 Operation principle

The basic behavior of pyro-electric materials entails the generation of charge when responding to heat flow following a temperature change, i.e. a pyroelectric sensor is a derivative or change responding device. A gradual charge disappearance occurs if the temperature stays at a constant level. When radiation on the sensor causes heating and thus an expansion of the lattice spacing, a change in electrical polarization will occur. So in a rigorous definition pyro-electricity can be understood as the temperature dependence of the spontaneous polarization $\vec{P}_S$ in the crystal, with $\vec{P}_S$ representing the dipole moment per unit volume of the material. The pyro-electric coefficient $\vec{p}$ is defined as:

$$\vec{p} = \frac{\partial \vec{P}_S}{\partial T}|_{E,X} \quad (7.180)$$

under constant elastic stress $\vec{X}$ and constant electric field $\vec{E}$ and it relates a change in temperature T to a change in electrical displacement $\vec{D}$ according to the relation $d\vec{D} = \vec{p} \cdot dT$. Thus the material becomes polarized under temperature changes resulting in the generation of a charge $\mathbf{Q}$ on the surface area A_s that may attract free charges from the environment (ions, electrons, dust) or, when electrodes are attached to this surface, may be measured as a current, the pyro-electric current I_p, generated to compensate the T-induced polarization charges, see figure (7.31). To pick up the charge $d\mathbf{Q}$ that is generated by the temperature change dT, the pyro-electric materials are fabricated typically in the shapes of a flat capacitor with area A with two electrodes on opposite sides and the pyroelectric material serving as a dielectric. The generated charge and the pyro-electric current I_p, under short-circuit conditions are then given by the equations:

$$d\mathbf{Q} = p(T) \cdot A_s \cdot dT \quad \text{and} \quad I_p(t) = p(T) \cdot A_s \cdot \frac{dT}{dt} \quad \Leftrightarrow \quad \bar{I}(j\omega) = j\omega \cdot p(T) \cdot A_s \cdot dT \quad (7.181)$$

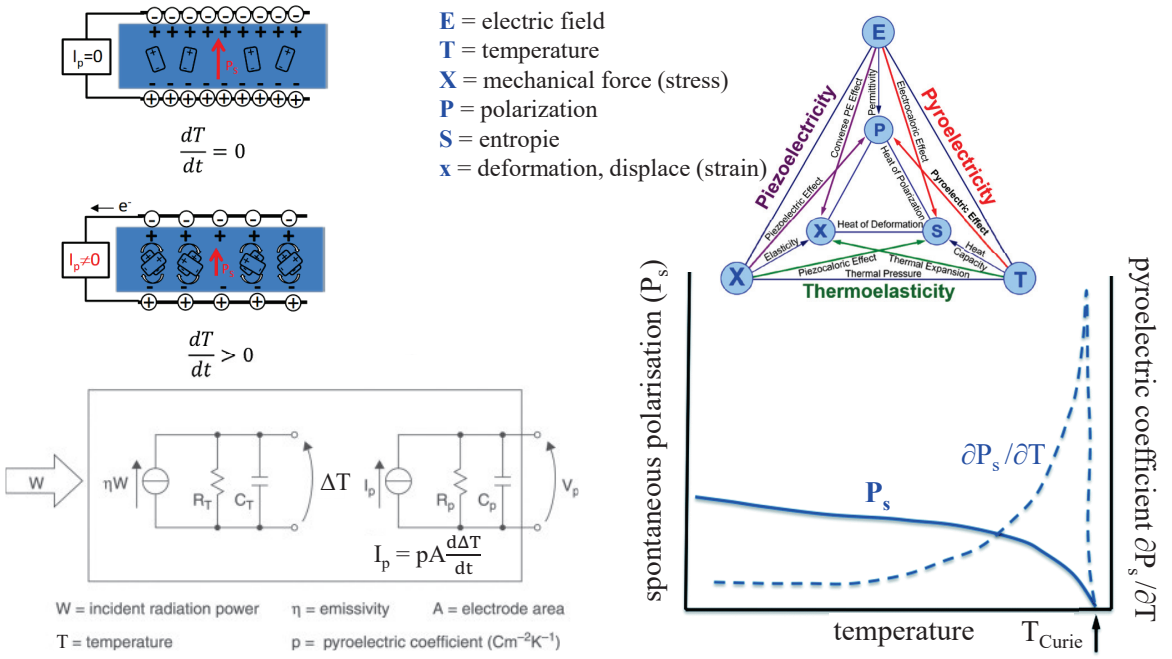


Figure 7.31: *Operational principle of a pyro-electric sensing element. (Upper right) Heckmann diagram depicting the various thermodynamic coupled electrical, mechanical, and thermal susceptibilities of crystal materials. Here, E, X, T, P, x, and S are the electric field, stress, temperature, polarization, strain, and entropy, respectively. In addition to the direct pyro-electric effect that relates polarization and temperature, other indirect thermodynamic paths are evident (e.g., pyro-electricity via temperature induced piezo-electricity). (Lower right) Spontaneous polarization* $\vec{P}_S$ *as a function of temperature ending at the Curie temperature* T_{Curie}*, the component of the pyroelectric coefficient normal to the electrodes* $p = \partial P_S/\partial T$ *is also shown. (Upper left) Illustrating the pyro-electric effect for a material with a given spontaneous polarization* $\vec{P}_S$ *(drawn in blue) made up of individual dipoles (crystalites, domains) sandwiched between electrodes (drawn in black). For a constant temperature and thus constant* $\vec{P}_S$ *the bound charges at the surface of the material are compensated by accumulated charges at the electrodes. For an increasing temperature* $\vec{P}_S$ *is prompted to decrease as the dipole moments, on average, diminish in magnitude. This decrease of* $\vec{P}_S$ *is pictured in the lower image by the horizontal tilting of the dipoles together with an increased amplitude of libration. A current flows from the electrodes to compensate for the change in bound charge that accumulates on the crystal edges.*

Here $p(T)$ is the component of the pyro-electric coefficient that results from the sum of the crystalline or molecular dipole moments oriented normal to the electrodes. Two contributions make up the pyro-electric effect: (i) the primary pyroelectric effect refers to a clamped crystal (meaning constant strain) for which a change in temperature induces directly a change in dielectric displacement due to the elongation or shortening of individual dipoles, also known as the *dimensional change*, and (ii) the secondary pyroelectric effect that results from crystal deformation and is mediated by the piezoelectric effect, the thermal expansion or contraction causes a strain in the crystal that

alters the dielectric displacement due to its piezo-electricity, which is also known as *electrostriction*. The temperature dependent libration of dipoles may also contribute to the secondary effect, especially in polymer pyroelectric materials. The sum of these effects, the so-called total pyro-electric effect under constant stress, is what is usually measured as p.
The most important application of the pyro-electric effect is the detection of long-wavelength infrared radiation, especially in the range from 8 to 14 μm, which covers the peak in emission of the black-body curve for objects at 300 K, that occurs at 10 μm. Consequently, IR sensors with high sensitivity at a wavelength of 10 μm can easily detect human beings and warm-blooded animals. Thermal sensors that convert heat into photons like pyro-electric detectors are cheap in manufacture, have a large spectral bandwidth and are sensitive over a wide temperature range. However, since they only react on temperature fluctuations, they are used mainly to detect moving objects in a scene or stationary objects with the aid of a light chopper. Some of the applications of IR pyro-electric sensors include intruder alarms and motion detectors for building protection, light control switches, radiometers, instant medical IR thermometers, flame and fire detectors, IR spectrometers, laser power meters, pollution monitors and thermal imaging systems.

7.8.2 Temporal frequency response and responsivity

The equivalent circuit of a pyro-electric sensor is also shown in the lower left part of figure (7.31). First we need to derive an expression for the temperature change $dT = \Delta T$. As in the case of the bolometer ΔT can be found from evaluating the heat balance equation: heat absorbed=[heat to warm the detector] - [heat losses] which is described in one dimension by the differential equation:

$$W_i(t) \;=\; C_T\frac{d(\Delta T)}{dt} + \frac{\Delta T}{R_T} + \sigma(T_s^4 - T_a^4), \text{ with } W_i(t) = \text{incident thermal power} \quad (7.182)$$

where R_T is the thermal resistance to the heat sink, C_T the heat capacity of the sensor, T_s the temperature of the sensor, T_a the ambient temperature and σ the constant for radiation losses. The incident thermal thermal power, if we model a chopped IR radiation beam, can be expressed as:

$$W_i(t) = \frac{W_i}{2}(e^{j\omega t} + 1) \quad (7.183)$$

Neglecting radiation losses the solution of this equation was already treated in case of the bolometer and is given by:

$$\Delta T == \frac{\eta W_i R_T e^{j\omega t}}{1 + j\omega R_T C_T} \;=\; \frac{\eta W_i R_T e^{j\omega t}}{1 + j\omega\tau_T}, \text{ with } \tau = R_T C_T \quad (7.184)$$

where τ_T represents the thermal time constant.
The output voltage of the sensor is the current $\bar{I}_p(j\omega)$ multiplied by the equivalent impedance of the pyro-electric sensor as represented by the equivalent circuit shown in figure (7.31):

$$\bar{V}_p = \bar{I}_p(j\omega)\frac{R_p(1/j\omega C_p)}{R_p + 1/j\omega C_p} \Rightarrow \frac{j\omega p(T)\eta W_i A_p R_T R_p}{(1 + j\omega\tau_T)(1 + j\omega\tau_p)} \quad (7.185)$$

The voltage responsivity can now be expressed as:

$$R_V = \frac{|\bar{V}_p|}{W_i} = \frac{\omega p(T)\eta A_p R_T R_p}{\sqrt{(1+\omega^2\tau_T^2)(1+\omega^2\tau_p^2)}} \tag{7.186}$$

where the responsivity is a combination of the thermal and the electrical characteristics

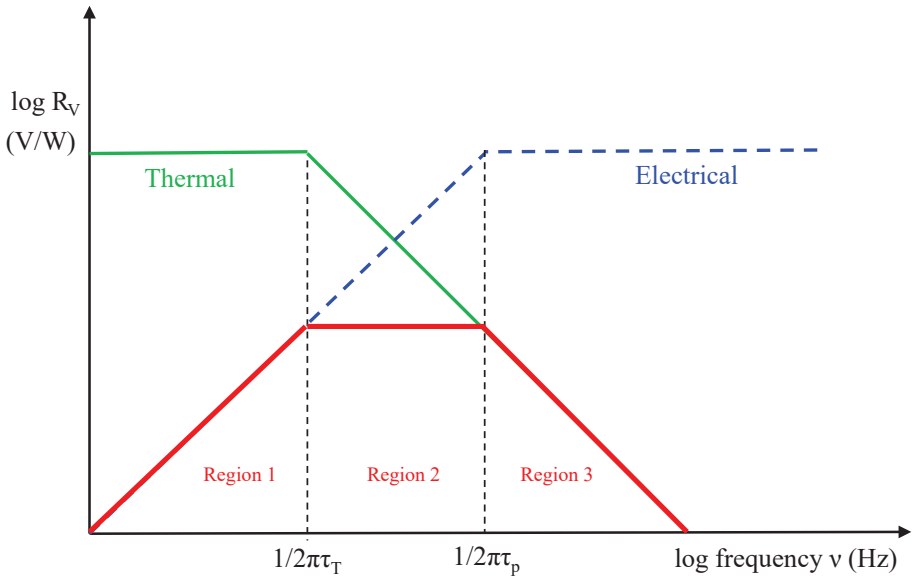


Figure 7.32: *Bode diagram of the voltage responsivity of a pyro-electric sensor element. The thermal (green) and electrical (blue) frequency responses are shown separately, the overall frequency response (red) is characterized by three distinct regions.*

of the pyrosensor. The sensor resistance R_p in equation (7.186) should actually be replaced by the parallel impedance of R_p and the load resistance R_L of the (pre)amplifier. Since the load resistance is much smaller than the sensor resistance R_p, this parallel impedance is actually almost equal to the load resistance, i.e. $\tau_p = R_L C_p$. A bode diagram can be developed for various values of R_L. If its value increases, the voltage responsivity increases but the temporal frequency response decreases. The Bode plot in Figure (7.32) shows three distinct regions: For $\nu < 1/(2\pi\tau_T)$ the voltage responsivity is proportional to frequency, in the region $1/(2\pi\tau_T) \leq \nu \leq 1/(2\pi\tau_p)$ we have compensating roll-off from the thermal and electrical filtering causing a constant responsivity and for $\nu > 1/(2\pi\tau_p)$ the responsivity is inversely proportional to frequency.

7.8.3 Defining noise sources

A total of four noise sources need to be considered for a pyro-electric sensor, i.e. phonon noise, thermal (Johnson) noise associated with the detector and the load resistor and the preamplifier noise. Note that there is no $1/f$ noise, since no bias current is flowing through the sensor.

- *Phonon noise*
  The noise associated with temperature fluctuations, i.e. phonon noise, was discussed in the section on bolometers. We derived expression (7.168) for the noise equivalent power NEP relevant here for region 2 of the voltage responsivity. Hence, if a sensor is connected to a heat sink via thermal conductance K_T at a temperature T, it will attain thermal equilibrium to establish zero mean power flow with a fluctuation expressed in equivalent thermal noise power:

$$NEP(\Delta\nu_N) = \sqrt{4kT^2 R_T \Delta\nu_N} \tag{7.187}$$

  This NEP can be related to a voltage fluctuation with:

$$\sqrt{\overline{\Delta V_T^2}} \;=\; NEP(\Delta\nu_N) \cdot R_V \tag{7.188}$$

  For a pyro-electric sensor this noise is rarely the limiting factor.

- *Thermal noise*
  Thermal noise is usually the dominant noise source for pyro-electric sensors. To evaluate this noise component, let us consider the sensor's parallel equivalent circuit that is displayed at the lower left of figure (7.31). One may actually interpret the capacitor C_p as a 'lossy capacitor', since ideally one would like the pyro-electric sensor to be a perfect capacitive element with no loss to a parallel resistor R_p, i.e $R_p = \infty$. This loss factor of a pyro-electric element can be used as a *figure of merit*, i.e. the lower the loss factor the better the sensor performance. The loss factor can be quantified with the aid of a *loss tangent*. This loss tangent of the capacitor C_p is defined as the tangent of the phase angle δ of the complex impedance $\bar{Z}_p = C_p//R_p$ (dividing the real component of $\bar{Z}_p$ by its complex component). In case of an perfect capacitor we would have $\tan\delta = 0$.
  The mean square thermal noise current associated with R_p is given by (see the earlier section on thermal noise):

$$\overline{\Delta I_{th}^2} \;=\; \frac{4kT\Delta\nu_N}{R_p} = \overline{\Delta I_{R_p}^2}(1+\omega^2 R_p^2 C_p^2) \Rightarrow \overline{\Delta I_{R_p}^2} = \frac{\overline{\Delta I_T^2}}{1+\omega^2 R_p^2 C_p^2} \tag{7.189}$$

  with $\Delta\nu_N$ the equivalent noise bandwidth and where we applied Kirchhoff's theorem to subdivide the total mean square noise current $\overline{\Delta I_T^2}$ between C_p and R_p. For the thermal mean square noise voltage we find:

$$\overline{\Delta V_{th}^2} = \overline{\Delta I_{R_p}^2} R_p^2 = \frac{\overline{\Delta I_T^2}}{1+\omega^2 R_p^2 C_p^2} R_p^2 = \frac{4kTR_p\Delta\nu_N}{1+\omega^2 R_p^2 C_p^2} \tag{7.190}$$

  We can now replace R_p by the figure of merit characterization $\tan\delta$ of the sensor that follows from:

$$\begin{aligned} \bar{Z} \;&=\; C_p//R_p = \frac{R_p}{1+j\omega R_p C_p} = \frac{R_p}{1+\omega^2 R_p^2 C_p^2} \;-\; j\frac{\omega R_p^2 C_p}{1+\omega^2 R_p^2 C_p^2} \Rightarrow \\ \tan\delta \;&=\; \frac{1}{\omega R_p C_p} \quad\Rightarrow\quad R_p = \frac{1}{\omega C_p(\tan\delta)} \end{aligned} \tag{7.191}$$

Substituting this expression for R_p in equation (7.190) we get for the thermal rms-noise-voltage of the pyroelectric element:

$$\sqrt{\overline{\Delta V_{th}^2}} = \sqrt{\frac{4kT\Delta\nu_N(\tan\delta)}{\omega C_p(\tan^2\delta\ +\ 1)}} \approx \sqrt{\frac{4kT\Delta\nu_N(\tan\delta)}{\omega C_p}},\ \text{since}\ \tan^2\delta \ll 1 \qquad (7.192)$$

The rms-noise-voltage associated with the load resistor R_L to be added in quadrature to $\overline{\Delta V_{th}^2}$ is simply given by:

$$\sqrt{\overline{\Delta V_{R_L}^2}} = \sqrt{\frac{4kTR_L\Delta\nu_N}{1+\omega^2R_L^2C^2}},\ \text{with}\ C\ \text{the total capacitance across}\ R_L \qquad (7.193)$$

- *Preamplifier noise*
  The choice of a preamplifier for a pyro-electric sensor is very critical. It is desirable to have the sensor loss-tangent-noise limited and consequently the preamplifier must have lower input noise than the intrinsic sensor noise. As discussed in earlier sections, the preamplifier has both current and voltage noise sources at its input. By employing high quality junction FET's (JFET), the input current noise source can be mostly eliminated,the typical input voltage noise levels of such high fidelity JFET's ranges between 1 and 5 nV $\text{Hz}^{-1/2}$.

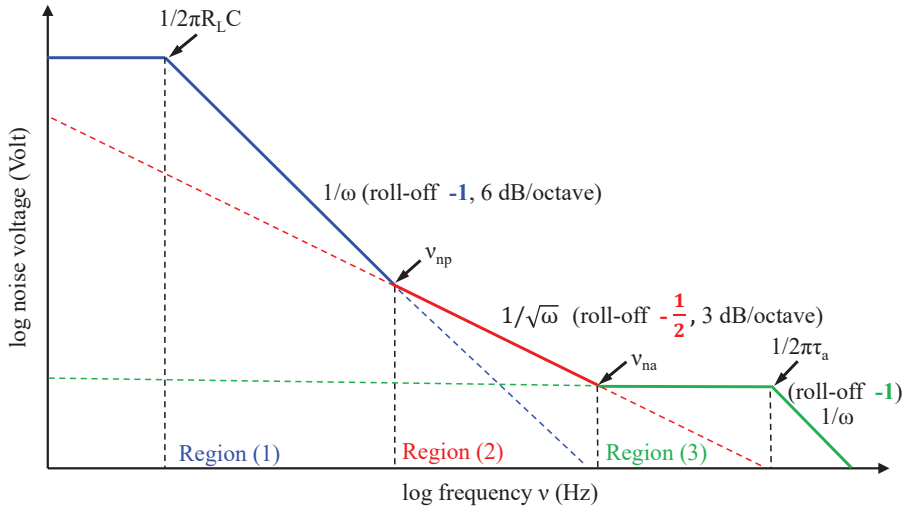


Figure 7.33: *Bode diagram of the dominating noise voltages across the frequency spectrum for a pyro-electric sensor. Three regions can be distinguished in which the load resistance, the sensor loss-tangent and the preamplifier noise input voltage dominate respectively.*

The total noise spectrum of a pyro-electric sensor can now be overviewed with a Bode plot of the various voltage noise components, see figure (7.33), in which the three distinct noise regions can be identified. In region (1), at the low frequency

end, the thermal noise of the load resistor R_L will dominate, see equation (7.193), which rolls off at $\nu_L = 1/(2\pi R_L C)$ with 6 dB/octave ($\propto 1/\omega$) up to frequency ν_{np}, where the load resistor noise equals the loss-tangent noise. As derived above, in this region (2) the sensor voltage noise expression varies as the square root of the frequency ($\propto 1/\sqrt{\omega}$) and rolls off at 3 dB/octave. Finally, at frequency ν_{na} the preamplifier noise and the loss-tangent noise become equal, so that the the noise for higher frequencies runs flat up to the high frequency cut-off at $\nu = 1/2\pi\tau_a$ of the preamplifier: region (3).
JFET's have an extremely large input impedance and very low gate-current leakage. Load resistors of the order of 10^{11} Ω are used for maximum responsivity. One problem with the sensor-preamplifier combination is its high sensitivity for microphonics due to the very large resistor values. Hence the sensor, the load resistor and the JFET should be placed in very close proximity to each other.

In summary, desirable characteristics of pyro-electric sensors are high Curie temperature, a large pyro-electric coefficient, large resistance value, a low loss-tangent and small heat capacity. The oldest material that showed pyro-electric behavior were crystals of the *mineral Tourmaline* which is actually non-ferro-electric, later on it was found that ceramic ferro-electric materials exhibit much larger pyro-electric coefficients. Pyro-electric sensors are particularly useful as uncooled sensors for infrared detection and are frequently applied in thermal imagers in the atmospheric MWIR band (3-5 μm) and the LWIR band between 8 and 14 μm (containing the 300 K peak emissivity). These thermal imagers contain microbolometer arrays that are based on ferro-electric (also pyro-electric) technology featuring *thin film Barium-Strontium-Titanate* or BST ($\mathbf{Ba_{1-x}Sr_xTiO_3}$ with *tunable* Ba/Sr fractions).

Chapter 8

Radiation sensors: image sensors

8.1 Charge Coupled Devices

8.1.1 Operation principle

A charge-coupled device (CCD) in its simplest form constitutes a closely spaced array of metal-insulator-semiconductor (MIS) capacitors. The most important among the MIS elements is the metal-oxide-semiconductor (MOS) capacitor, made from silicon (Si) and employing silicon dioxide (SiO_2) as the insulator. The charge position in the MOS-array of capacitors is electrostatically controlled by voltage levels. Dynamical application of these voltage levels and their relative phases serves a dual purpose: injected charges (e.g. due to radiation generated electron-hole pairs) can be stored in the MOS capacitor and the built-up charge (i.e. charge packet) can subsequently be transferred across the semiconductor substrate in a controlled manner. Monolithic CCD arrays are available for imaging in the near-infrared, the visible and the X-ray range, for the mid-infrared (up to 20 μm), hybrid fabrication is employed in which the infrared-sensitive detector material is sandwiched to a CCD array of cells for read-out.

In the following sections a short description is given of the storage and transfer mechanisms and of a few focal-plane CCD-architectures.

8.1.2 Charge storage in a CCD

Two basic types of charge coupled structures exist. In the one type, charge packets are stored very close to the interface between the semi-conductor (Si) and the overlaying insulator (SiO_2). Since the charges reside virtually at the surface of the semiconductor layer, these devices have become known as Surface channel CCDs (SCCDs). In the other type, charge packets are stored at some distance away from the surface of the semiconductor, such devices have become known as Bulk or Buried channel CCDs (BCCDs).

Figure (8.1) shows a single CCD-electrode comprising a metal gate, separated by a thin oxide layer (few times 0.1 μm) from a p-type semiconductor (hole-conduction). The designation *metal gate* is used to imply an electrode with very high conductivity, in

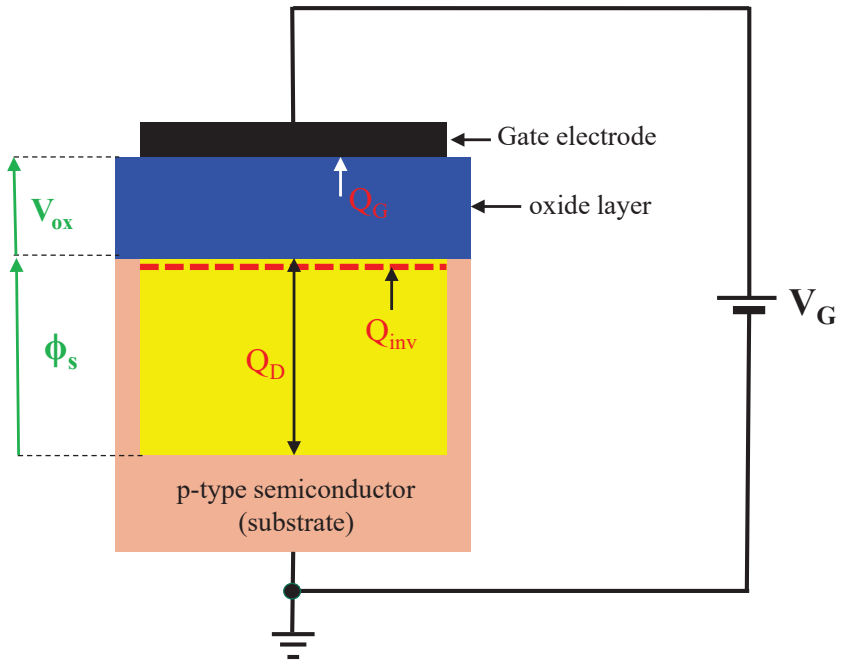


Figure 8.1: *Single CCD-electrode. Q_G is positive, Q_{inv} and Q_D are negative.*

the CCD practice *highly doped polycrystalline silicon* is used instead of metal, so-called *poly-silicon gates* or in short *polygates*, however we shall keep the metal gate designation for clarity.

Prior to the application of a voltage bias to the gate electrode, there will be a uniform distribution of holes (majority free charge carriers) in the p-type semiconductor. As the gate electrode is made positive however, the holes are repelled immediately beneath the gate and a so-called depletion layer (devoid of free charge) is created. Increasing the gate voltage causes the depletion region to extend further into the semiconductor and the potential at the semiconductor/insulator interface (ϕ_s) becomes increasingly positive. Eventually, a gate voltage bias is reached at which the surface potential ϕ_s becomes so positive that electrons (i.e. the minority charge carriers in the p-type material !) are attracted to the surface where they form an extremely thin(≈ 0.01 μm thick), but very dense (in terms of charge density) layer, a so-called *inversion layer*.

Basically the electrons reside in the deep potential well at the semiconductor surface and do not recombine with holes, since the latter are immediately repelled from the depletion layer upon creation. Therefore, if radiation is incident on a single CCD electrode, the electrons from the produced electron-hole pairs will be stored in the inversion layer, the holes will be repelled from the depletion region. The electrostatics of the CCD(MIS) cell shown in figure (8.1) has been treated when discussing the working principle of a MIS element in *inversion mode*. The surface potential of the depletion layer V_{ds}, in short now designated ϕ_s, of the CCD(MIS) cell comprising a p-type substrate is given by (see for derivation equation (7.127) in Chapter 7):

$$V_{ds} = \phi_s = \left(V_G + \frac{Q_{inv}}{C_{ox}}\right) + V_0 - \left[2V_0\left(V_G + \frac{Q_{inv}}{C_{ox}}\right) + V_0^2\right]^{\frac{1}{2}}, \qquad (8.1)$$

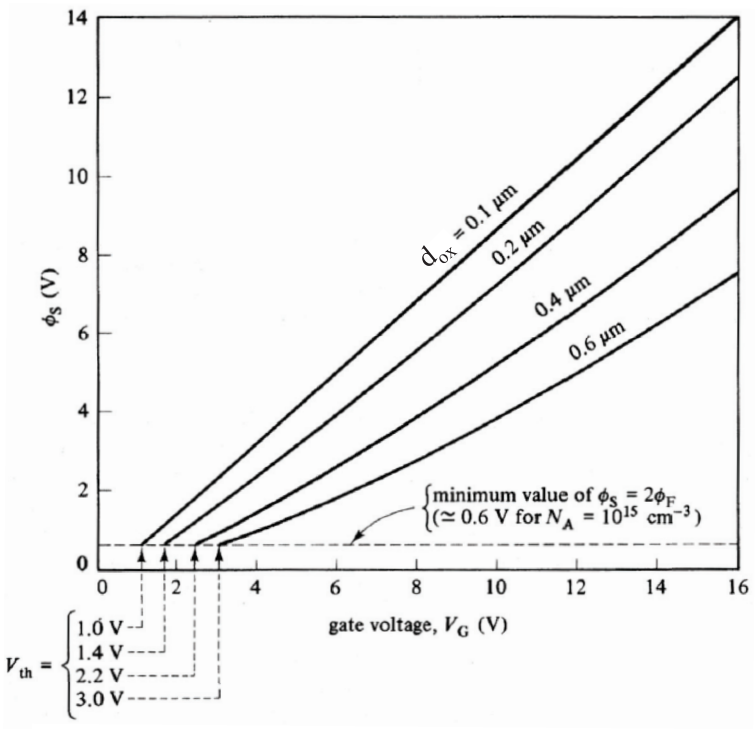


Figure 8.2: *Variations of the surface potential ϕ_s with the gate voltage V_G for different values of the oxide thickness d_{ox}. The charge in the inversion layer Q_{inv} is zero in all cases. Figure taken from Beynon & Lamb (1980).*

The value of ϕ_s as a function of V_G (with $Q_{inv} = 0$) and as a function of Q_{inv} (with V_G = 10 and 15 V respectively) is displayed in figure 8.2 and figure 8.3 for a few thicknesses of the insulating oxide layer.

The surface potential ϕ_s can be interpreted as a *potential well*, the depth of which is determined by the magnitude of the inversion charge packet Q_{inv} (in analogy with a bucket partially filled with fluid). From the figures it can be seen that ϕ_s is practically a linear function of Q_{inv} and V_G, this arises from the fact that the constant V_0 is relatively small compared to typical values for V_G; i.e. $V_0 \approx 0.14$ Volt for a 0.1 μm thick oxide layer. Since V_G = 10 - 15 Volt, we have in *good approximation*:

$$\phi_s = V_G + \frac{Q_{inv}}{C_{ox}} \tag{8.2}$$

Note: in the above computation the difference in workfunction between the metal and the semiconductor has been neglected, i.e. the so-called *flat-band condition* has been assumed. In practice this can always be achieved by adaptation of the potential V_G. A formula for the magnitude of this adaptation was derived in the section on MIS elements in Chapter 7. Moreover ϕ_s has a minimum value related to the Fermi-level in the semi-conductor, this boundary condition has been omitted from the above formulae, since it is of no direct relevance to the storage principle.

Storage of charge at the Si-SiO_2 interface introduces potential losses of accumulated

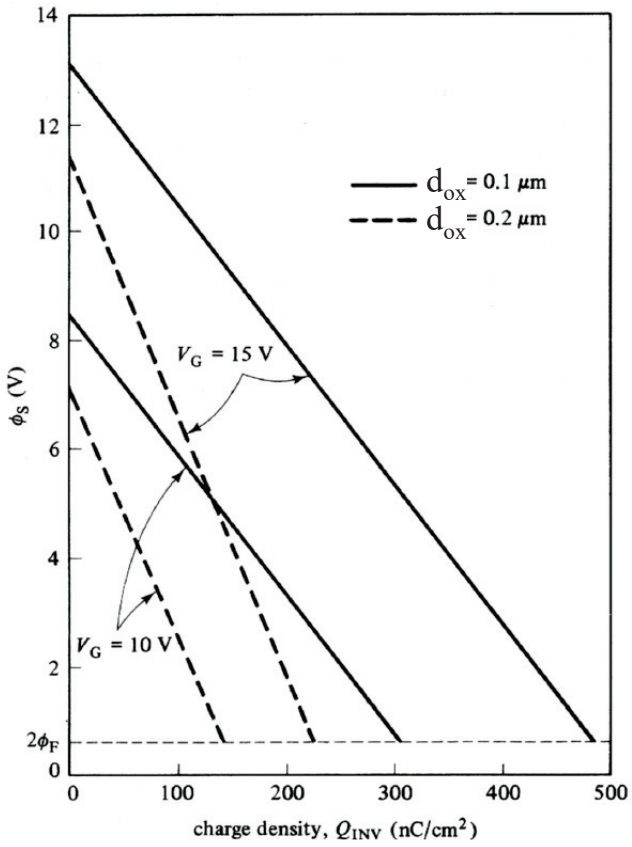


Figure 8.3: *Variations of the surface potential ϕ_s as a function of the inversion charge Q_{inv} for two values of the gate voltage V_G and of the oxide thickness d_{ox}. Figure taken from Beynon & Lamb (1980).*

charge during charge transport, owing to charge trapping in the atomic surface states. To prevent such losses the charge packet should preferably be kept at a potential minimum separate from the Si-SiO_2 interface, i.e. a minimum in the bulk-silicon has to be generated.

The realization of such a potential minimum, detached from the interface plane of the semiconductor with the oxide is the subject of the following paragraph in which we shall discuss the principle of the *buried channel* CCD (BCCD). The structure of such a BCCD cell is shown in figure (8.4). The BCCD element comprises a relatively lightly doped semiconductor substrate (in this case p-type) on top of which a much thinner, more heavily doped, layer of n-type material is deposited (thickness of a few μm). An insulating oxide layer and a metal electrode are placed on top of the semiconductor pn-junction and are similar to those in the SCCD case. A heavily doped n^+ contact is embedded in the n-type layer in order to apply a positive bias voltage to the pn-junction, a so-called *contact diffusion*. The gate electrode can be held at zero potential or biased at a positive voltage, whereas the p-substrate is grounded. This structure gives rise to the development of two depletion layers:

- one depletion layer below the gate electrode with a thickness x_1.

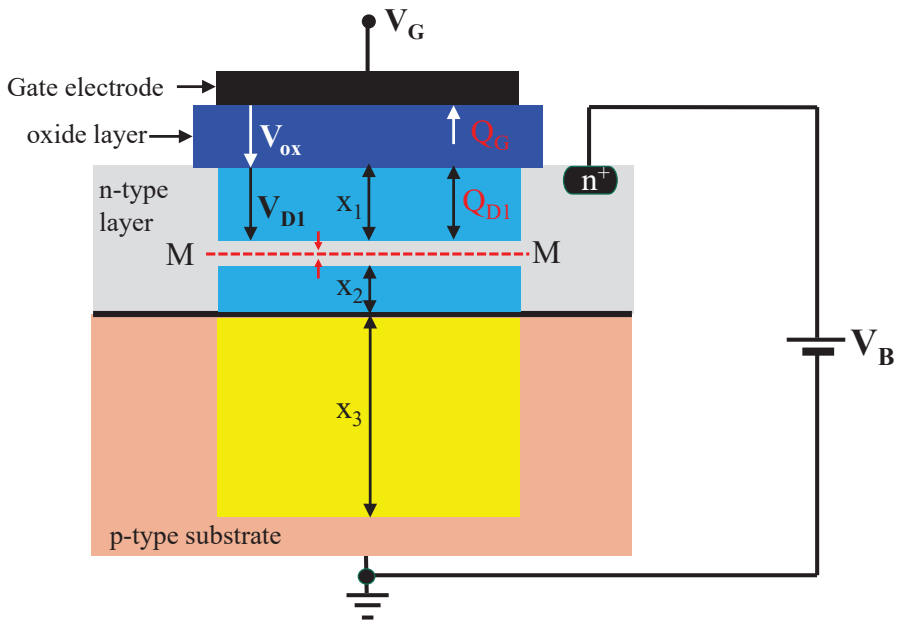


Figure 8.4: *Single BCCD-electrode.* Q_G *and* Q_{inv} *are negative charge surface densities,* Q_{D1} *is a positive charge surface density.*

- a second depletion layer around the reverse-biased pn-junction, with thickness $x_2 + x_3$ in the n-type (x_2) and the p-type (x_3) material respectively.
If the reverse bias voltage V_B is sufficiently high, the two depletion layers will touch in the n-type material at position M, this is the so-called pinch-off voltage. For V_B values larger than pinch-off, the potential at position M, V_M, becomes independent of V_B. In practical cases the pinch-off voltages vary between 5 and 20 Volts. Above the pinch-off voltage, the potential at location M is the maximum potential within the n-type layer below the metal gate contact. Hence, electrons that are produced or injected in this structure will be collected at this position of *minimum electron potential* and are subsequently stored in a plane through M parallel to the boundary plane with the oxide.
Pay attention however: the charge storage mechanism in a BCCD is fundamentally different from that in a SCCD. In the latter the electrons are the *minority charge carriers* of the semiconductor type under consideration that are stored in an extremely thin inversion layer with a very high charge volume density. In a BCCD, on the contrary, the electrons that are stored are the *majority charge carriers* of the relevant semiconductor material, they substitute for part of the electron population that were originally removed by the electric field in the depletion layer (complete depletion of the n-type layer took place above pinch-off). In fact, the collected 'replacement' electrons do produce a thin layer of *neutral semiconductor material* centered around the plane through location M. This neutral layer will gradually increase in width if it becomes filled with more charge.
Application of a voltage bias V_G at the gate electrode allows the possibility to influence the potential distribution in the semiconductor. In order to have field induced depletion in the upper n-type layer, the condition $V_G < V_B + V_{FB}$ needs to be met, i.e. the gate potential should be negative relative to the surface potential of the n-type layer to

initiate the formation of depletion layer x_1. In case $V_G > V_B + V_{FB}$ no depletion but accumulation will occur (see section on MIS elements).
First, we shall now derive a relation between V_G and V_M without any free charge in the buried channel. Next we shall treat the case that free charge carriers are released and stored in the channel. Assuming that the flat-band condition has been fulfilled (e.g. $V_{FB} = 0$), we can deduce the following relations for the potentials from figure (8.4):

$$V_M = V_G + V_{ox} + V_{D1} \tag{8.3}$$

With a surface charge density Q_G on the gate electrode, an opposite surface charge density Q_{D1} in the n-type layer and an electric field strength $\vec{E} = \vec{0}$ beyond the gate electrode and in the plane through M (potential maximum) we have:

$$\begin{aligned} Q_G + Q_{D1} &= 0, \text{ with } V_{ox} = -\frac{Q_G}{C_{ox}} \quad \Rightarrow \quad V_{ox} = \frac{Q_{D1}}{C_{ox}} = \frac{qN_d x_1}{C_{ox}} \qquad (8.4) \\ V_{D1} &= \frac{qN_d x_1^2}{2\epsilon_0\epsilon_s} \text{ (from the theory of depletion layers, } N_d = \text{donor concentration)} \\ &\Rightarrow \; V_M = V_G + \frac{qN_d x_1}{C_{ox}} + \frac{qN_d x_1^2}{2\epsilon_0\epsilon_s} \qquad (8.5) \end{aligned}$$

Moreover, from the theory of the pn-junction (see Chapter 4) we also have:

$$V_M + \Phi_{pn} = \frac{qN_d(N_d + N_a)}{2\epsilon_0\epsilon_s N_a} x_2^2 \tag{8.6}$$

In practice Φ_{pn} is ≈ 0.5 Volt and $\ll V_M$, therefore we shall neglect (omit) this built-in-potential factor in the rest of the derivation.
The 'pinch-off' condition dictates that $x_n = x_1 + x_2$, with x_n the thickness of the n-type layer. Substitution of x_2 in equation (8.6) and elimination of V_M with the aid of relations (8.5) and (8.6) yields a solution for x_1 and subsequently for $x_2 = x_n - x_1$.We can then derive the following relation between V_M and V_G:

$$\begin{aligned} V_M &= \eta \left\{ x_n - \xi + \left[\xi^2 - \frac{2\epsilon_0\epsilon_s(\eta x_n^2 - V_G)}{(2\eta\epsilon_0\epsilon_s - qN_d)} \right]^{\frac{1}{2}} \right\}^2 \quad \text{with} \qquad (8.7) \\ \xi &= \epsilon_0\epsilon_s \left[\frac{(2\eta x_n C_{ox} + qN_d)}{C_{ox}(2\eta\epsilon_0\epsilon_s - qN_d)} \right] \quad \text{and} \quad \eta = \left[\frac{qN_d(N_d + N_a}{2\epsilon_0\epsilon_s N_a} \right] \qquad (8.8) \end{aligned}$$

Figure (8.5) shows V_M as a function of the gate voltage voor specific combinations of values regarding the thickness of the oxide layer (d_{ox}), the thickness of the n-type layer (x_n) and the impurity concentrations N_d and N_a of both semiconductor types. In all cases no free charges are present in the buried channel.
Let's now analyze the situation where free charges are injected into an empty BCCD channel. This charge will produce a broadening of the channel around the plane through M, the width of the channel is determined by the amount of injected charge Q_i: $Q_i = qN_d x_M$. Suppose that from the total charge Q_i we find a part Q_{i1} at the gate side of

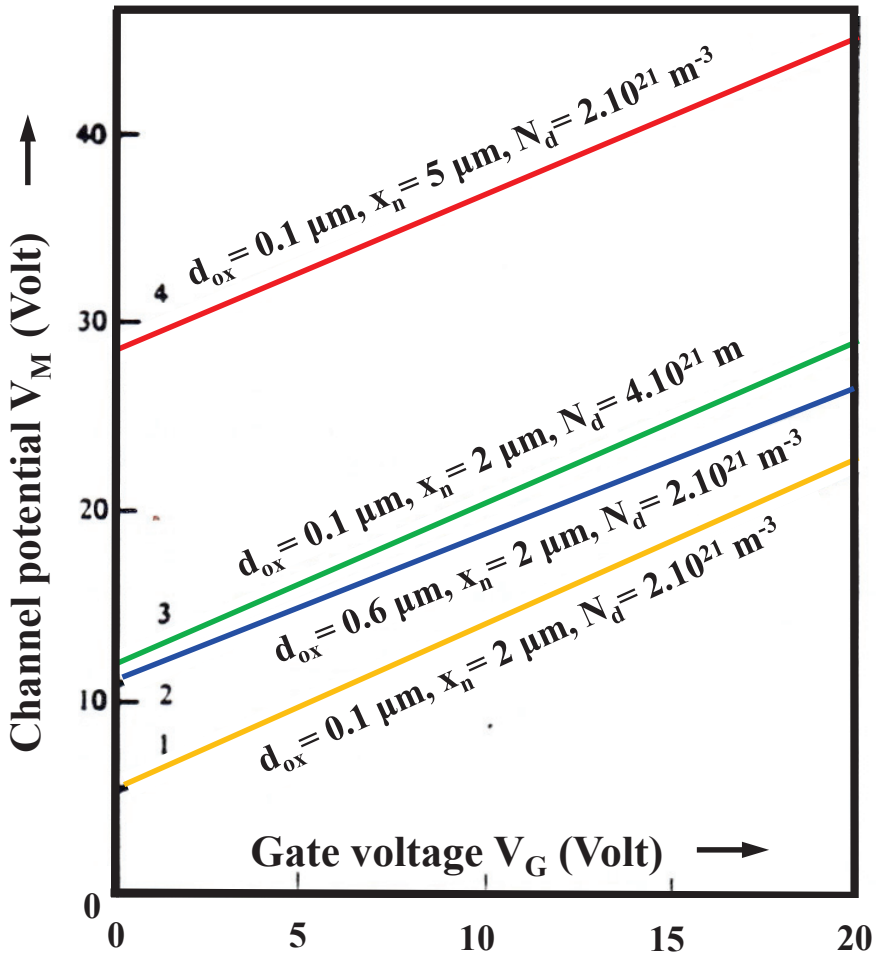


Figure 8.5: *Variation of the channel potential V_M for various buried channel structures fabricated on a p-type substrate with an impurity concentration $N_a = 10^{20}$ m^{-3}. No free charges are present in the channels.*

M with a thickness x_{M1} and at the side of the pn-junction we find the complementary part Q_{i2} with a thickness x_{M2}. The following relations now hold:

$$\begin{aligned} Q_i &= Q_{i1} + Q_{i2} \\ x_M &= x_{M1} + x_{M2}, \quad \text{and the thicknesses of the depletion regions reduce to:} \\ x_1^* &= x_1 - x_{M1} \\ x_2^* &= x_2 - x_{M2} \end{aligned}$$

For this case equations (8.5) and (8.6) still hold as long as we replace the entities x_1 and x_2 by $x_1^*$and x_2^*. Taking $x_0 = x_n - x_M$, the solutions for x_1^* and V_M^* can then be

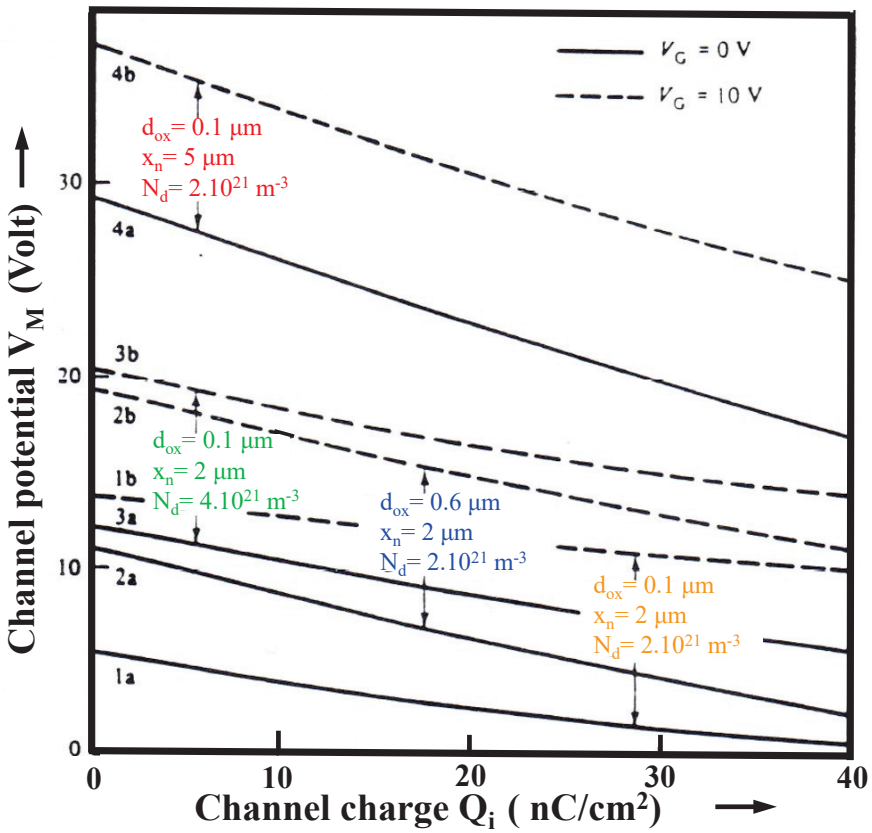


Figure 8.6: *Variation of channel potential V_M with surface charge density Q_i for the same BCCD structures as in figure (8.5) when gate voltages of 0 Volt (solid lines) and 10 Volt (dashed lines) are applied.*

written as:

$$x_1^* = \xi^* - \left[\xi^{*2} - \frac{2\epsilon_0\epsilon_s(\eta x_0^2 - V_G)}{(2\eta\epsilon_0\epsilon_s - qN_d)}\right]^{\frac{1}{2}} \tag{8.9}$$

$$V_M^* = \eta\left\{x_0 - \xi^* + \left[\xi^{*2} - \frac{2\epsilon_0\epsilon_s(\eta x_0^2 - V_G)}{(2\eta\epsilon_0\epsilon_s - qN_d)}\right]^{\frac{1}{2}}\right\}^2 \quad \text{with} \tag{8.10}$$

$$\xi^* = \epsilon_0\epsilon_s\left[\frac{(2\eta x_0 C_{ox} + qN_d)}{C_{ox}(2\eta\epsilon_0\epsilon_s - qN_d)}\right] \tag{8.11}$$

Figure (8.6) shows the channel potential V_M as a function of the injected charge Q_i for the same combination of parameter values as displayed in figure (8.5) using the same color for identification. These plots confirm what intuitively is expected. If the amount of charge in the channel increases, the potential drops. BCCD structures with a thicker oxide layer or a thicker n-layer show a larger potential drop per unit charge than the thinner structures, since the latter possess larger capacitance. Compared to the SCCD, the dependence on the structure parameter values is similar, except that the charge density and consequently the charge storage capacitance is lower in the BCCD case as compared to the SCCD. The reason for this is that the charge in a BCCD is positioned

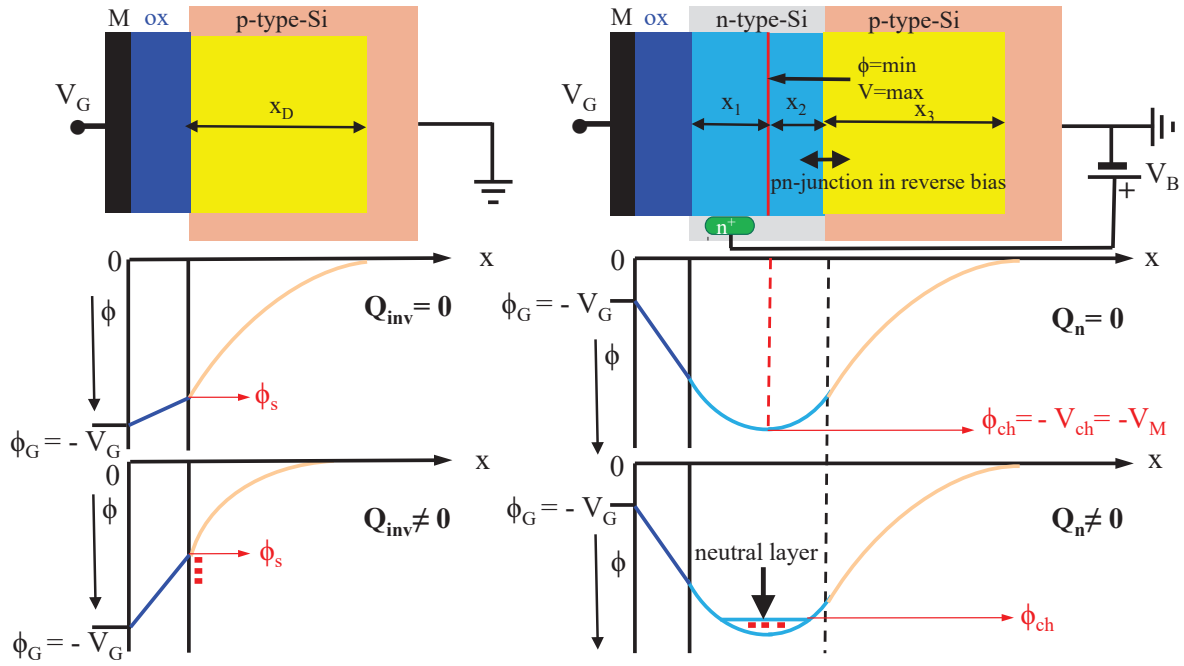


Figure 8.7: *Surface and channel electron potentials for a SCCD (left) and a BCCD (right), with empty and partly filled wells. The creation of a field induced depletion layer by the gate electrode in the n-Si layer imposes the condition* $V_G < V_B + V_{FB}$*, with* V_B *the reverse bias voltage on the pn-junction and* V_{FB} *the flat band voltage.*

a few μm below the gate, the effective gate-to-channel capacitance is therefore much lower than for the case of the SCCD, where the charge resides at only a fraction of a μm below the gate electrode. This difference is partly compensated by the value of ϵ_s for Si, which is much smaller than for SiO_2 with a ratio 12:4.
Figure 8.7 shows also the comparison between an empty and partly filled SCCD, (a) and (b), and an empty and partly filled BCCD, (c) and (d). As discussed above, in the latter case the charge packet flattens the minimum of the potential well, this can then be regarded as a 'neutral layer' of width x_M.

8.1.3 Charge transport in a CCD

The way in which a charge packet is transported through a CCD structure (charge-coupling) is illustrated in figure 8.8. Assume the charge packet is stored initially in the potential well under the second electrode (from the left in the figure) which is biased to say 10 V. All CCD electrodes need to be kept above a certain threshold voltage (V_{th}) to ensure that each MOS capacitor operates in the inversion-mode. In this example a minimum required bias voltage of 2 Volt was selected, assuming $V_{th} < 2$ Volt. The potential well under the 10 V electrode is much deeper than those under the 2V electrodes and, provided it is not "overfilled" with charge, it is only in this well that charge will be stored. Suppose that the bias on the third electrode is now increased to 10 V. If the second and third electrode are sufficiently close, both wells will merge. The equations that govern the charge transfer process comprise the current density and the

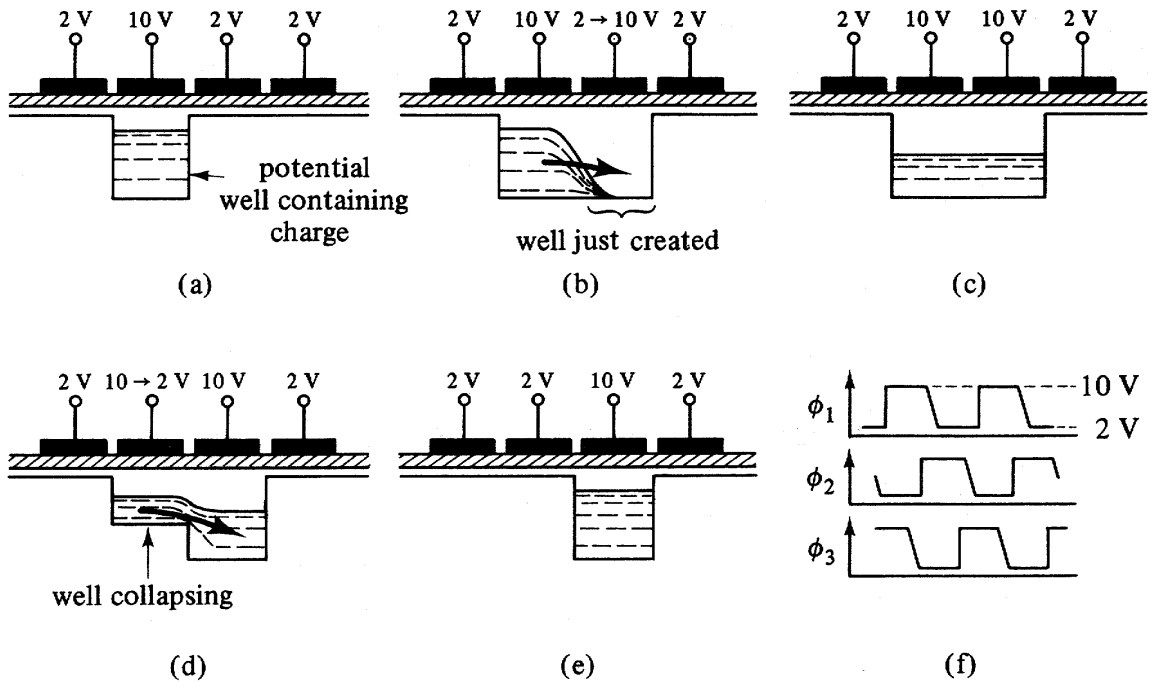


Figure 8.8: *(a)-(e) Transport of a charge packet in a CCD. (f) clocking waveforms for a three phase CCD. Figure taken from Beynon & Lamb (1980).*

continuity equation. For an n-type channel (electron charge) these equations are:

$$j(x,t) = q\mu_n n(x,t)\frac{\partial\phi(x,t)}{\partial x} + qD_n\frac{\partial n(x,t)}{\partial x} \tag{8.12}$$

$$\frac{\partial n(x,t)}{\partial t} = \frac{1}{q}\frac{\partial j(x,t)}{\partial x} \tag{8.13}$$

where charge propagation is in the x direction (no vector treatment), $n(x,t)$ is the electron density, μ_n the electron mobility and D_n the electron diffusion coefficient. This diffusion coefficient is related to the mobility of the charge carriers μ_n through

$$\frac{D_n}{\mu_n} = \frac{kT}{q}, \tag{8.14}$$

the Einstein relation that we derived when treating the physical principle of the pn-junction in Chapter 4 equations (4.19) and (4.20).
The first right-hand term in equation 8.12 refers to drift of the charge packet under a gradient in electric potential, the second term refers to thermal diffusion under the presence of an density gradient. The drift component arises from two causes. The first cause derives from *self-induced fields* in the presence of a gradient inv charge concentration, charges of the same type will mutually repel and reshuffle the concentration of charge carriers in such a way that this gradient becomes zero. Secondly, due to so-called *fringing fields*, in which charges are forced to move due to the existence of electric fields generated by the various voltage levels on the gates.
The equalization of charge concentration under electrodes two and three described above

is hence governed by the drift speed due to self-induced fields and by thermal diffusion. In practice these two processes are frequently described by introducing an effective diffusion coefficient $D_{n_{ef}}$, since both arise from the existence of a gradient in carrier concentration. For large charge packets (large concentration difference), the process is mainly determined by the self-induced drift, for small charge packets (e.g. $< 1\%$ of the well charge capacity) thermal diffusion dominates and slows the transfer process down. Returning now to figure 8.8, if the charge is evenly distributed under electrodes 2 and 3 (situation c), reducing the potential on the second electrode to 2 V will introduce a fringe field and the remaining content of this well will drift into the third well owing to the fringe field. The final situation shown in (e) is a deep potential well and its associated charge packet moved one position to the right. By applying a succession of varying voltages to the gate electrodes, the charge packet can be transported through the CCD structure in a controlled manner. The gate electrodes can be grouped together with each group or *phase* connected to a different voltage clock. Various CCD structures and electrode systems are employed, the one described above requires a three-phase clocking system and is therefore referred to as a *three phase CCD*. Each image pixel requires in this case three gate electrodes to store and transport a single pixel charge packet. In addition three independent, mutually properly phased, clock lines are necessary to secure unambiguous transport of each charge packet.

It is obvious that the transfer mechanisms described are not perfect, i.e. not all of the charge is transferred and some signal charge is left behind. The charge transfer efficiency (CTE) is the ratio of charge transferred to the initial charge present. A complementary term often used in the literature on these devices is the transfer inefficiency (CTI), i.e. the charge lost to the initial charge packet: CTI = 1 - CTE. Typical values of CTE are of the order 0.99999 for a good device or CTI = 10^{-5}. As already mentioned before the CTE of a SCCD is, without special measures not discussed here, considerably worse compared to the BCCD due to the surface interface trapping states. Moreover, due to the relaxation of charges from the traps, the so-called trapping noise in a SCCD is at least an order of magnitude larger than in BCCDs, which explains the preference for BCCDs in certain applications. Till now, only charge transport along the gate electrodes in one direction had been discussed. It is of course necessary to limit the extent of the potential well in the orthogonal direction. This *lateral confinement* of the charge packet is achieved by a so-called *channel-stop diffusion*, which comprises a heavily doped region of the semiconductor relative to its neighboring regions. This region exhibits a large conductivity σ relative to the surrounding material and quenches the surfaces potential ϕ_S so that no depletion region can be formed. In this way one dimensional columns (or rows) are implemented in the CCD structure along which the charge transfer occurs, isolated from its neighboring columns (rows). At the positions of the channel stop diffusion (p^+ or n^+ material), the field oxide separating it from the gate structure is much thicker as to prevent break-down of the p^+ (n^+) contact to the conducting gate electrodes. A three dimensional display of a BCCD structure is shown in figure 8.9.

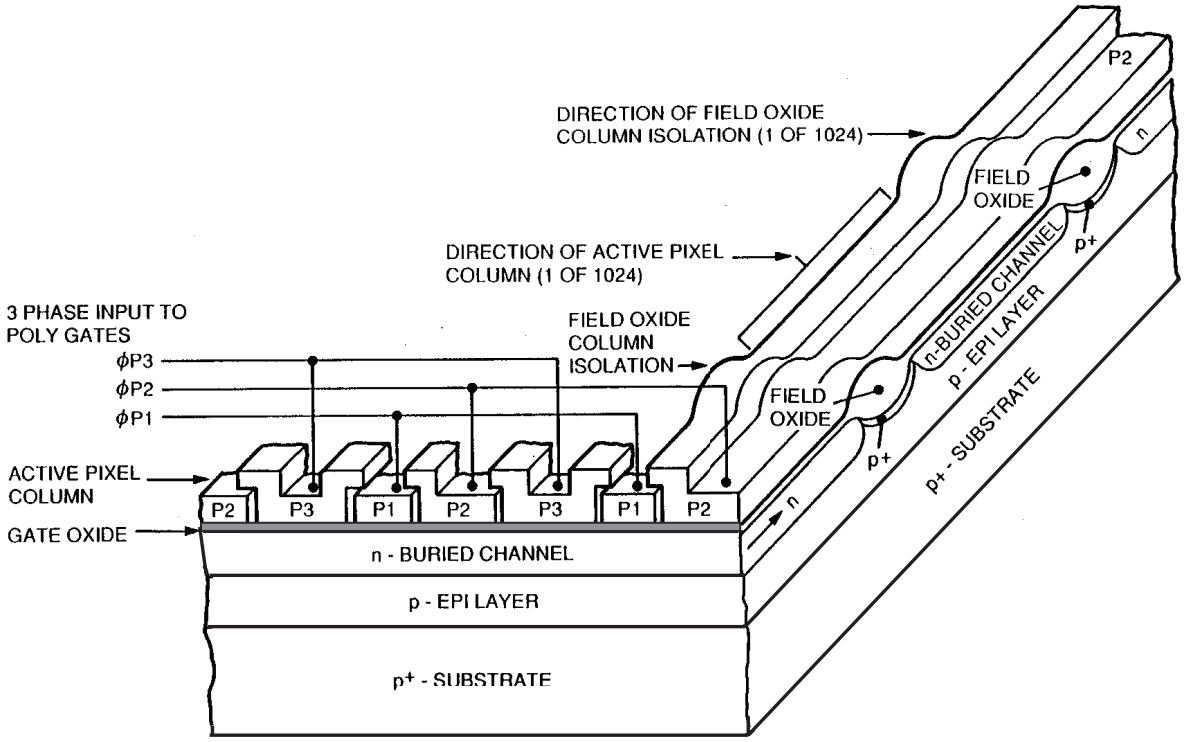


Figure 8.9: *Three dimensional display of a BCCD structure.*

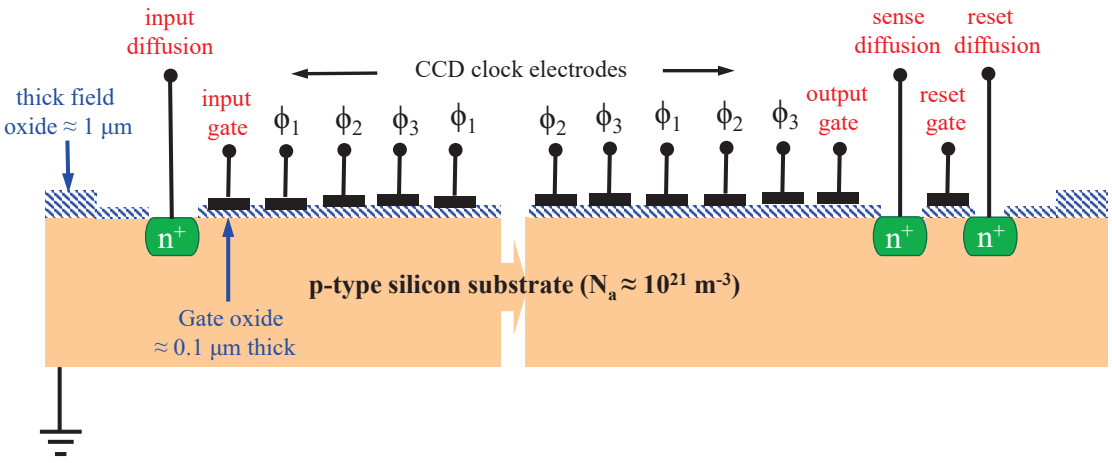


Figure 8.10: *Diagram of a complete 3-phase CCD system, including input and output terminals, the n^+ represent heavily doped n-type contact regions.*

8.1.4 Charge injection and processing

In case the CCD is intended to be used as an analogue or digital shift register, an input source of minority charge carriers is required. This can be accomplished with the aid of strongly doped piece of n-type material, a so called n^+ input diffusion. Figure (8.10)

shows this in a diagram that displays a complete 3-phase CCD system with all the input and output terminals indicated. These minority charge carriers can then be injected into the CCD structure via an input gate. If the CCD is used as an image sensor, such an input arrangement is not necessary, the minority charge carriers are generated by an incident radiation beam of photons that releases charge carriers within 1 μm of the semiconductor layer. During the light integration period, the clock pulse train is put on hold at a specific phase, that causes for instance a deep potential well under clock terminal ϕ_1 that can be filled with charge with a speed proportional to the incident light intensity. After a preselected integration period the accumulated charge can be shifted out by the register. Arriving at the output of the register, the charge can be detected with an output diode, the sense diffusion, that is operated at a high reverse-bias and acts as a current sink. This current can be measured as a voltage over a load resistor. However, mostly a transimpedance amplifier is integrated on the CCD chip to minimize the capacitive load. The easiest solution involves a MOSFET with its gate directly coupled to the sense diffusion. A reset gate coupled to a reset diffusion, that is kept at the highest positive potential, takes care that after passage of each charge packet through the sense diffusion the latter is reset by activating (i.e. pulsing) of the reset gate.

Instead of a three-phase structure, a two-phase structure can also be implemented.

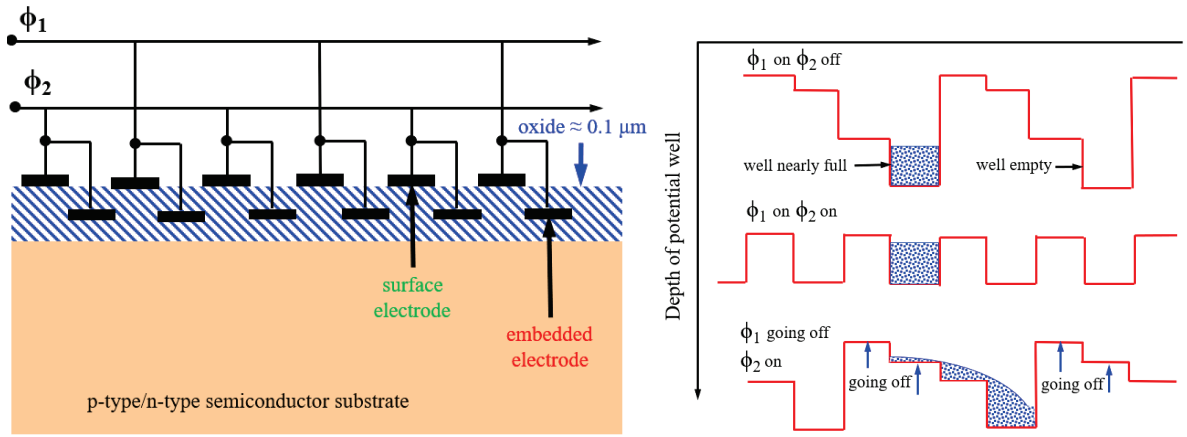


Figure 8.11: *Gate structure for a two-phase CCD and the associated profiles of potential wells during the clocking sequence.*

This solution incorporates two groups of electrodes that are both separated from the semiconductor surface with two different thicknesses of the oxide layer. In this way two different depths of potential wells are created for each phase of the clock signal. The shift process operates as follows, see figure (8.11):

phase ϕ_1 *on, phase* ϕ_2 *off* $\Rightarrow$ charge resident in the deepest well under ϕ_1.
phase ϕ_1 *on, phase* ϕ_2 *on* $\Rightarrow$ no charge transport, contrary to a three-phase CCD.
phase ϕ_1 *off, phase* ϕ_2 *on* $\Rightarrow$ the total charge flows to the deepest well under ϕ_2.

Motion to the left is prohibited since the potential barrier under the thick oxide always remains too high.

By chosing the amplitudes of the clock pulses on the ϕ_1 and ϕ_2 electrodes differently it

even becomes possible to operate a two-phase CCD in a 'one-clock-phase mode'. Two-phase CCD's allow for a simpler lay-out of a CCD-chip and simpler operation based on only two clock lines. In contrast to the three-phase CCD, where the charge packets can be transported in either direction the two-phase CCD only allows unidirectional transport. In practice, this constitutes only rarely a problem.

8.1.5 Charge capacity and transfer speed in CCD structures

In general, when designing a CCD, a compromise needs to be met between the desirability of a large signal (= charge) to obtain a good signal to noise ratio, and the necessity to avoid excessively large cell sizes and clock voltages that are needed for the storage of large charge packets. So the amount of charge that can be stored in a CCD pixel depends mainly on two variables: the clock voltages and the electrode area in combination with a particular SCCD or a BCCD structure.
Regarding a SCCD, the full-well storage capacity follows from:

$$\mathbf{Q}_{ox} + \mathbf{Q}_{inv} + \mathbf{Q}_{dep} = 0 \Rightarrow -\mathbf{Q}_{inv} = \mathbf{Q}_{ox} + \mathbf{Q}_{dep} = A_{el}C_{ox}(V_G - \phi_s) - A_{el}(2qN_a\epsilon_0\epsilon_s\phi_s)^{\frac{1}{2}} \tag{8.15}$$

with A_{el} the area of the gate electrode and $\mathbf{Q}_{inv}$ the *total inversion charge* in Coulomb's (pay attention: *not* the charge surface density Q!). For a completely filled potential well we have $\phi_s \approx 2E_F/q \approx 0.6$ Volts. Taking characteristic values for the above parameters, N_a =10^{21} m^{-3}, $d_{ox} = 0.1\,\mu$m, A =10x20 μm and V_G =10 Volt yields:

$$\mathbf{Q}_{inv} = -(0.63 - 0.028)\,\text{pC} \Rightarrow\approx -0.6\,\text{pC} \Rightarrow 3.7 \cdot 10^6\,\text{electrons} \tag{8.16}$$

Apparently the second term in equation (8.15) is quite small in comparison to the first term and $\phi_s \ll V_G$, we can then write in good approximation:

$$A_{el}C_{ox} = \frac{\mathbf{Q}_{inv}}{V_G} \quad \Rightarrow \quad \mathbf{Q}_{inv} = A_{el}C_{ox}V_G \tag{8.17}$$

This result shows that a good estimate of the charge capacity can be obtained by treating the the CCD structure as a mere solid state capacitance.
For a BCCD we have for the total signal charge:

$$\mathbf{Q}_{sig} = -A_{el}qN_dx_M \tag{8.18}$$

For x_M we already derived a somewhat complicated formula in the section on CCD-structures and, hence, a simple expression for the charge storage capacity, like in the case of the SCCD, is not readily available. The expression is also more complicated due to the finite depth of the charge storage and transport in the Si. However, eventually, the ratio in charge handling capacity between SCCD and BCCD structures can be shown to approximate:

$$\frac{\mathbf{Q}_{SCCD}}{\mathbf{Q}_{BCCD}} = 1 + \frac{\epsilon_{ox}x_{max}}{2\epsilon_{Si}d_{ox}} \tag{8.19}$$

in which x_{max} represents the maximum width of the neutral layer in the BCCD when the full storage capacity is reached. For typical values of $x_{max} = 2\ \mu$m and $d_{ox} = 0.1$

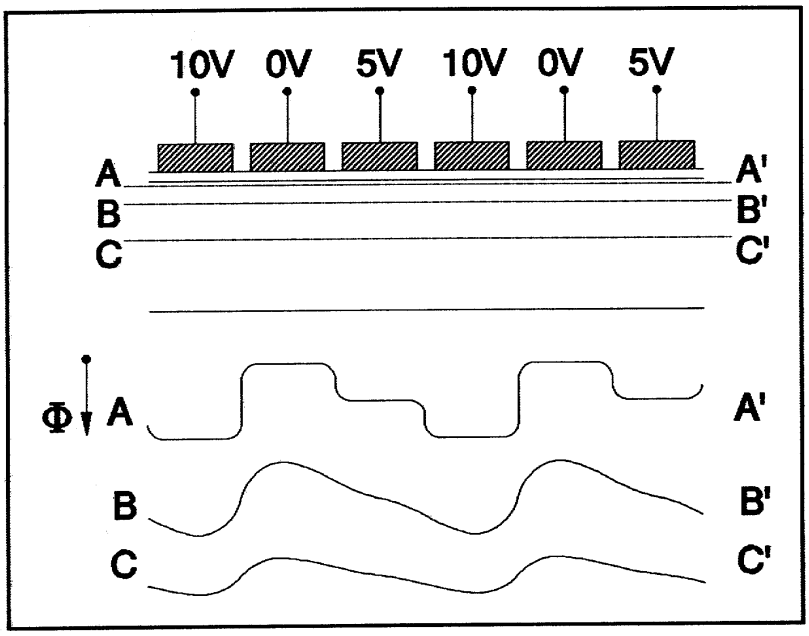


Figure 8.12: *Fringing fields at the Si-SiO_2 interface at several depths. Figure taken from Beynon & Lamb (1980).*

μm, this ratio amounts to about 3. Consequently, creation of an embedded channel lowers the charge handling capacity by a substantial factor.
The intrinsic speed of charge transport in a CCD structure is governed by transport equation 7.46, depending on the time constants for self-induced drift, thermal diffusion and fringe field drift. In a SCCD the time constant for self-induced drift is a function of charge density, C_{ox} and the inter-electrode spacing. For $C_{ox} = 1$ pF, $Q_{inv} = 10^{16}$ elementary charges m^{-2} and spacing 25 μm, the time constant $\tau_{Si} = 0.14$ μs. The time constant for thermal diffusion of an electron packet ($D_n \approx 10$ $cm^2 \cdot s^{-1}$) amounts to 0.25 μs. On this basis the high frequency limit would appear to be a few MHz, however the fringing field of the neighboring gate electrodes aid the transfer considerably, especially when thermal diffusion is dominant and clocking frequencies up to 15 MHz can be used for SCCDs. In the case of the BCCD the speed is potentially dominated by the fringing field of the neighboring gate due to the depth of the charge channel. The potential levels do not exhibit the flat structures encountered at the Si-SiO_2 interface as displayed in figure 8.8 but possess a continuous gradient along which charge can drift swiftly to a potential minimum. The situation is schematically shown in figure 8.12 for a few depths of the charge channel. BCCDs, with usually smaller charge packets, can therefore be read out at much higher frequencies, up to 300 MHz, while still retaining acceptable values of the CTE for several imaging applications.

8.1.6 Focal plane architectures

The two-dimensional CCD imaging arrays can be subdivided into *frame-transfer* (FT) imagers, *interline-transfer* (IL) imagers and *frame-interline-transfer* (FIT) devices. The

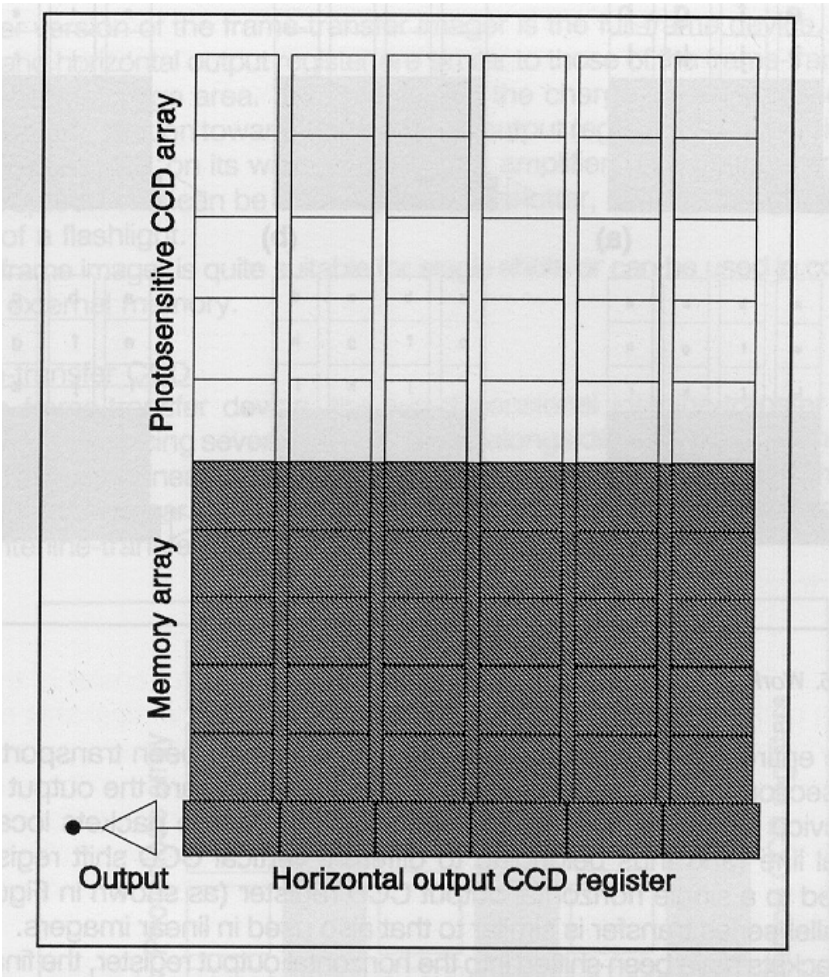


Figure 8.13: *Architecture of a frame-transfer image sensor. Figure taken from Theuwissen (1995).*

FT and IL architectures are briefly described below. Figure 8.13 shows the device architecture of a frame-transfer image sensor, the operation is shown in figure 8.14. The CCD comprises a photosensitive array and a memory array coupled to a linear output register. The light sensitive top part of the device, or image section, integrates the charges produced by the impinging photons over a predefined integration time. At the end of this period the charge packets are transferred along the columns into the memory array (figure 8.14(b)). This transfer needs to be done quickly to prevent disturbance by light falling on the image section during read-out. Subsequently each row in the memory section is clocked into the linear output register from which it is shifted to the output stage amplifier (video signal). During transport of all video information from the storage area to the output stage, all CCD cells in the image array are again biasedin the integration mode, allowing the build-up of the next image.
A simpler version of the frame-transfer imager is the *full-frame* device, which lacks the storage section. During read-out the incoming light should be interrupted to avoid disturbances. This can be done by using a shutter which cuts out the light path to the imager. The typical architecture of the interline-transfer imager is displayed in

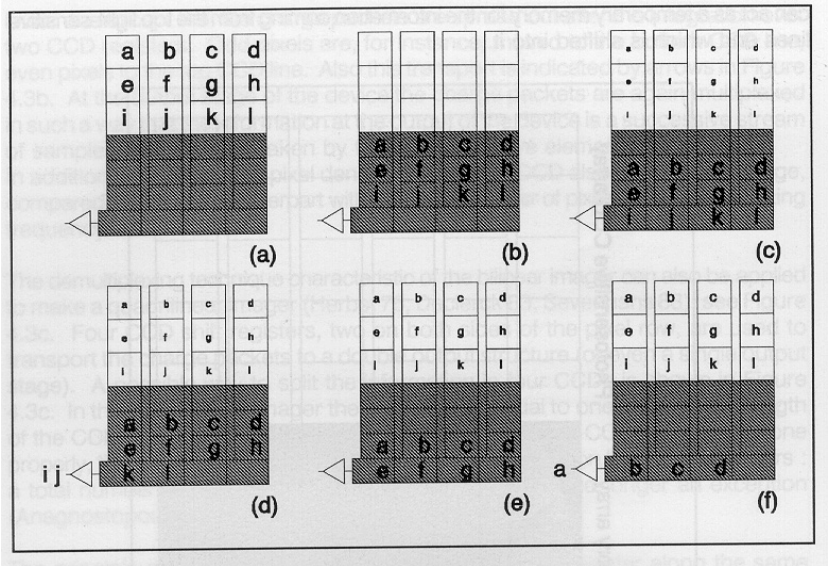


Figure 8.14: *Working principle of a frame-transfer image sensor. Figure taken from Theuwissen (1995).*

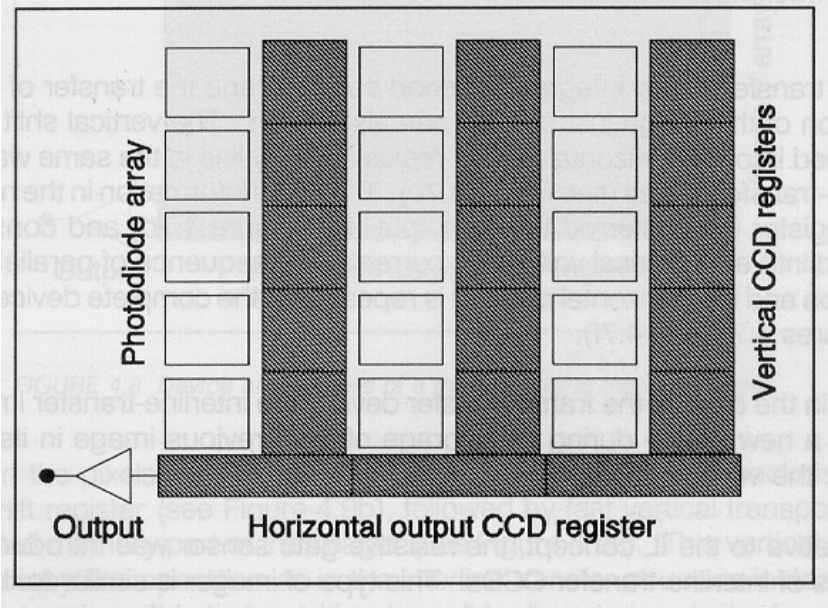


Figure 8.15: *Architecture of an interline-transfer image sensor. Figure taken from Theuwissen (1995).*

figure 8.15. In this case the photo-sensitive pixels are located near the shielded CCD transport register. The integration of charge takes place in the photo-sensitive pixels during exposure to the radiation source, at the end of the integration time the charge packets from the photo-sensitive columns are transferred in parallel to the vertical registers alongside them, see figure 8.16(b). The vertical shift registers are emptied into the horizontal-output register row by row which are read-out serially at the output. The advantage of this architecture is the very short transfer time to the parallel storage

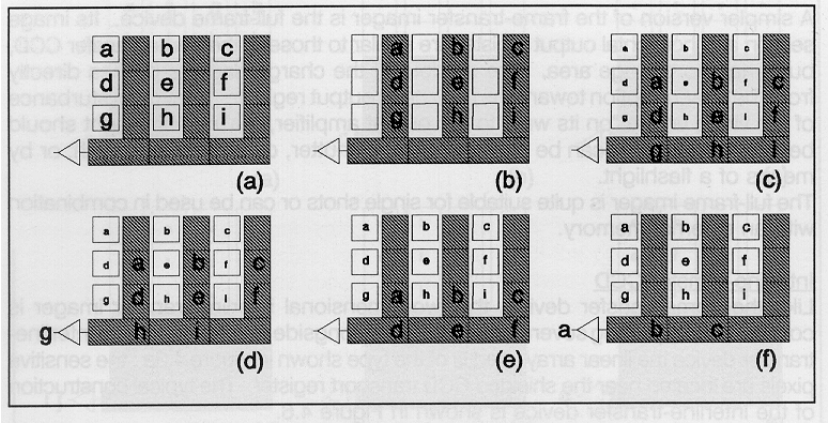


Figure 8.16: *Working principle of an interline-transfer image sensor. Figure taken from Theuwissen (1995).*

column as compared to the serial shift to the storage section in the case of the frame-transfer CCD. On the other hand, the presence of periodic storage columns in the image area may introduce severe aliasing effects close to the spatial Nyquist frequency of the CCD-array.
If a pixel has a characteristic size Δx, the sampling of an image can be described with an array of normalized window functions: $\frac{1}{\Delta x}\Pi\left(\frac{x}{\Delta x}\right)$. The spatial frequencies (s) associated with this window function is obtained from its Fourier transform: $\mathrm{sinc}s\Delta x$. If the pitch of the pixels in the image plane is given by x_0, the Nyquist frequency follows from $s_N = \frac{1}{2x_0}$. Normalizing the spatial frequency s on s_N yields the geometrical MTF for a linear array of pixels Δx with pitch x_0 ($s_0 = s/s_N$):

$$MTF_{geo} = \mathrm{sinc}\frac{s_0 \Delta x}{2x_0} \tag{8.20}$$

For contiguous pixels, like in a frame-transfer CCD array, $\Delta x \approx x_0$ yielding:

$$MTF_{FT} = \mathrm{sinc}\frac{s_0}{2} \tag{8.21}$$

with $MTF_{FT} = 0$ for $s_0 = 2$, i.e. $s = 2s_N$. For non-contiguous pixels, like an interline transfer CCD array,$\Delta x \approx \frac{x_0}{2}$ yielding:

$$MTF_{IL} = \mathrm{sinc}\frac{s_0}{4} \tag{8.22}$$

with $MTF_{IL} = 0$ for $s_0 = 4$, i.e. $s = 4s_N$.
This analysis indeed shows that the MTF of the interline-transfer CCD extends twice as high above the Nyquist frequency as compared to the frame-transfer CCD, giving rise to potential Moiré fringes due to aliasing. This effect can only be avoided by limiting the spatial frequencies of the input-image signal to s_N.

8.1.7 Wavelength response of CCDs

For visual light applications, two possibilities can be considered to irradiate the CCD, i.e. irradiation through the front surface and irradiation through the back-surface, so-

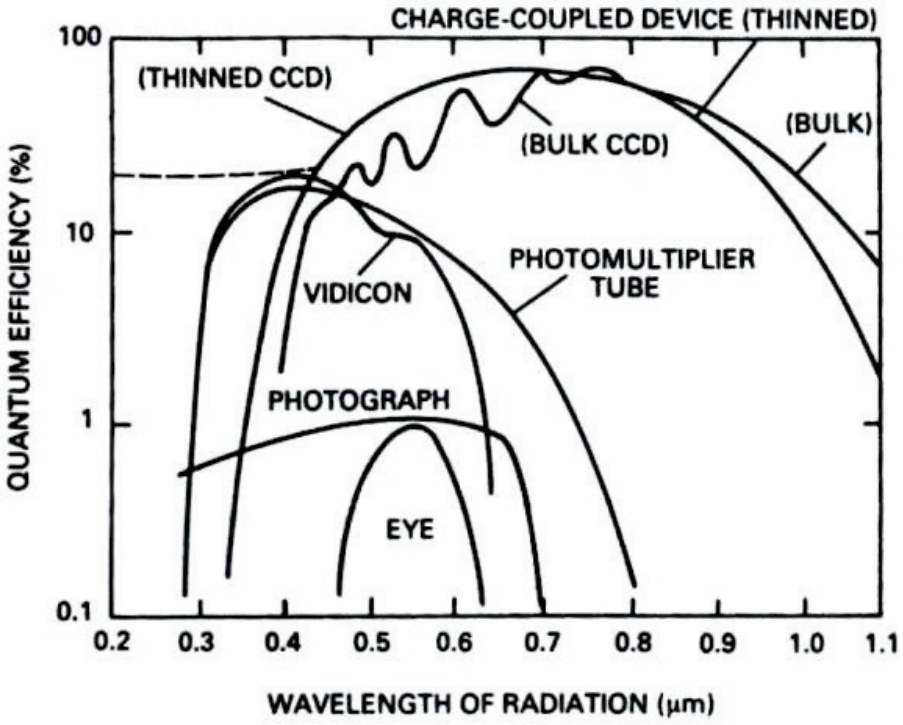


Figure 8.17: *CCD spectral response (i.e. quantum efficiency) in comparison to the human eye, a photographic plate, a photomultiplier tube and a vidicon.*

called back-illumination. Illumination through the front surface, i.e. the side which contains the gate structure, requires the usage of poly-silicon gate electrodes which transmit light. A problem encountered in this case entails the strongly wavelength dependent absorption and interference effects occurring in the thin poly-silicon gate layer (≈ 0.5 μm) and the thin oxide layer ($\approx 0.1 - 0.2$ μm). The current responsivity and the interference effects are shown in figure 8.17. Also, the blue-responsivity is strongly suppressed by absorption in the poly-silicon gate material. Back-illumination requires thinning of the Si-substrate to ensure that the photon-generated charge reaches the potential wells. The charge transport can be aided by building an electric field gradient into the semi-conductor, this electric field gradient can be provided by increasing the substrate doping concentration in the regions close to the silicon surface. This increased doping has the effect of *bending* the energy bands at the back of the surface so as to accelerate the photon-generated carriers towards the front surface and the potential wells. This technique is particularly useful for increasing the blue-responsivity where the charge carriers are generated close to the rear silicon surface. This back-face response can be further improved by minimizing the reflection of light from the back surface employing a $\lambda/4$ thick layer of silicon monoxide at the wavelength of interest. Figure 8.17 shows a quantum efficiency of about 50 % with back-illumination, which can be raised to a peak efficiency of about 90% by using the proper antireflection coating. Figure 8.18 shows a wafer with an array of CCDs for optical light with different pixel sizes, architectures and array sizes.

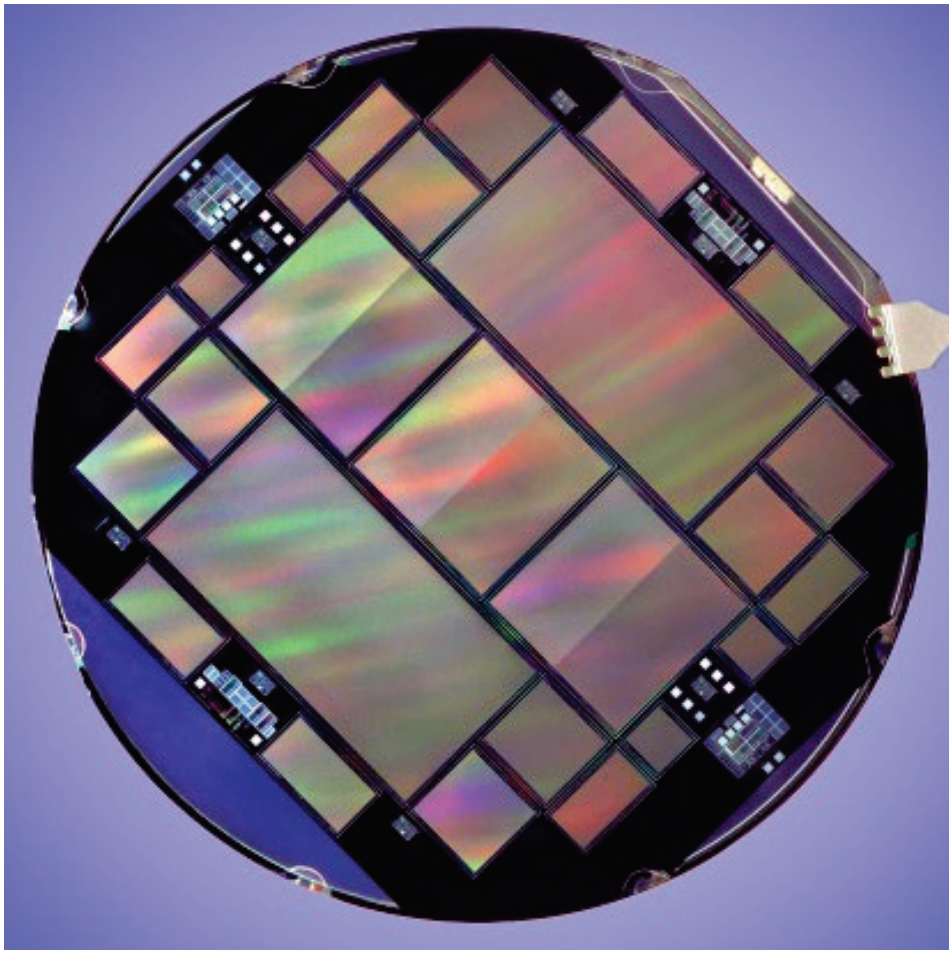


Figure 8.18: *The two large CCD's are 2048 x 4096 15 μm^2 devices for astronomy. The three central CCD's are 25202 12 μm^2 and 28802 10.5 μm^2 prototypes for the SNAP (Super-Nova Acceleration Probe) satellite camera. The wafer also contains near-square 15 μm^2, 12 μm^2, 10.5 μm^2, and 9 μm^2 CCD's, 1024 x 512 15 μm^2 CCD's for radiation testing, and 1200 x 600 15 μm^2 devices with different 2-stage amplifier designs. In addition there are test diodes and a number of other monitoring devices (from Lawrence Berkeley National Laboratories).*

Non-visible imaging with CCDs is normally divided into three wavelength ranges:

Wavelengths longer than 1 μm, for which the photo-electric absorption coefficient in silicon is too low ($h\nu_{IR} <$ bandgap) and all photons pass through the structure without being absorbed. For this reason infrared photons need to be converted first into electrons, e.g. by means of so-called Schottky-barrier structures in which pixels are used made out of platinum-silicide (PtSi). An array of these detectors is then coupled to a CCD read-out system.

The responsivity in the thermal IR can theoretically be extended in this way to approximately 5.6 μm.

For wavelengths shorter that 0.4 μm and longer than 10 nm the opposite is the case as compared to the IR range: the absorption is very high since not only silicon, but also

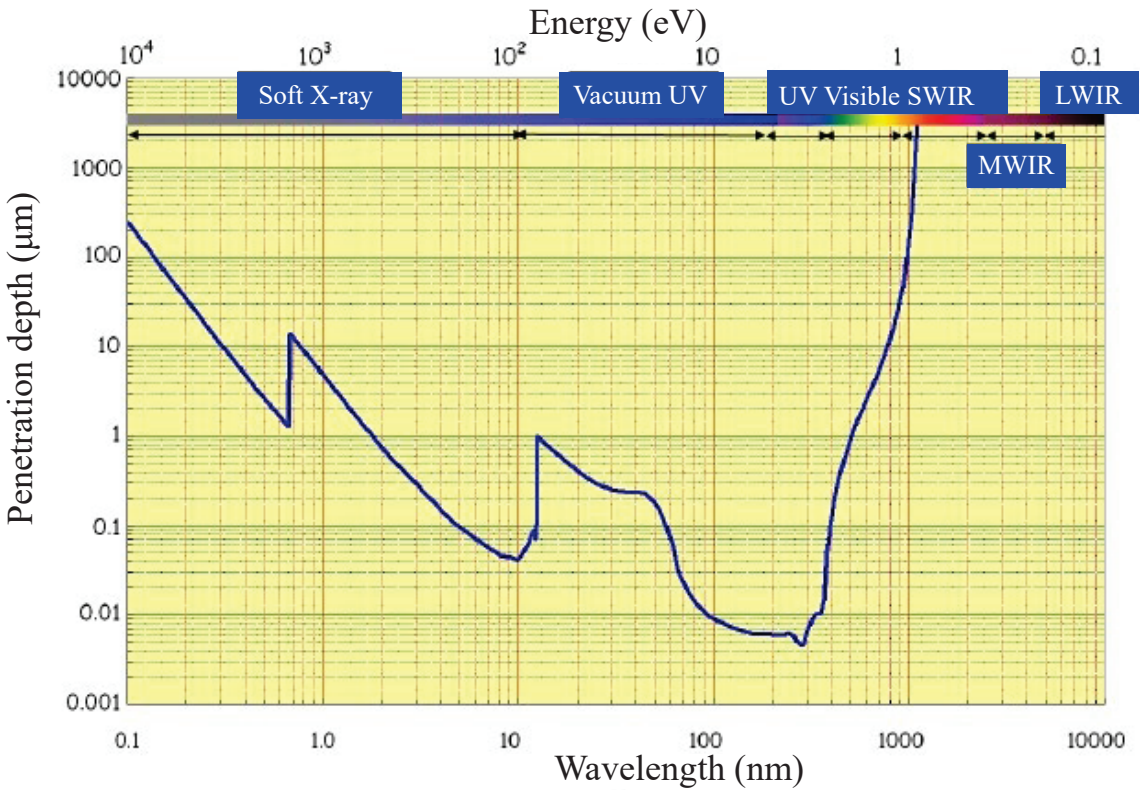


Figure 8.19: *Penetration depth of silicon as a function of wavelength from far-infrared (LWIR) to X-rays.*

the silicon oxide layer has a very high absorption coefficient in this wavelength regime. In figure 8.19 the penetration depth in silicon is displayed over a wide wavelength range, it is clear that the UV photon will be stopped in the top layers above the CCD. To avoid this problem several solutions are possible:

Deposition of an UV-sensitive phosphor on top of the active area of the imager which down-converts the energy of the UV-photons to longer wavelengths. An example is Coronene, which fluoresces in the green portion of the visible spectrum under UV exposure.

Also, as in the visible regime, back-illumination is an option. However, due to the high absorption, the substrate of the CCD has to be thinned to typically 10 μm, which is a very expensive process and requires subsequently very delicate handling of the device. An alternative is deep depletion of a lightly-doped, high-resistivity substrate, such that the depletion region under the CCD gates extends to the back of the silicon wafer. The charge carriers generated by the UV illumination are swept to the front side into the potential wells by the electric field of the deep depletion layer. This approach does not require the extreme thinning mentioned above for conventional high-doped material, a thickness of 50 μm still allows adequate collection of charge owing to the deep depletion field.

8.1.8 Noise sources in CCD's

It is convenient to subdivide the many sources of noise within a CCD in three categories:

- Noise arising from injection of charge into the device.
- Noise attributed to the fluctuations in the charge transferred from one gate electrode to the next.
- Noise introduced by the charge sensing circuitry.

In the following sections we shall now discuss the most important noise sources within these three categories.

Input noise

Practically all schemes that are used for the electrical injection of signals into a CCD structure utilize a MOS transistor (MOST). The input circuitry can therefore be modeled by an equivalent injection channel resistor R_i that connects at the interface between the oxide and the depletion layer to the capacitance C_{dep} of the depletion layer, in addition it 'sees' in parallel the capacitance of the oxide layer C_{ox} that includes the contributions from stray capacitance in the lay-out architecture. So we have a total input capacitance $C_i = C_{dep}//C_{ox}$, in contrast to the situation as seen from the gate electrodes where C_{dep} and C_{ox} are 'seen' in series (see also the treatment of the MIS element). The noise equivalent circuit is shown in figure (8.20), with $\overline{V_{th}^2}$, representing the quadratic thermal noise voltage produced by the injection channel resistor R_i, equal to $4kTR_i\Delta\nu_c$, where $\Delta\nu_c$ is the frequency bandwidth of the system. Our interest is now the assessment of the rms-fluctuation in the number of electrons on the capacitor C_i. We can write the (one sided) power spectral spectral density of the noise voltage on C_i as:

$$S_{C_i}(\omega) = 4kTR_i \left| \frac{1/j\omega C_i}{R_i + 1/j\omega C_i} \right|^2 = \frac{4kTR_i}{1 + \omega^2 C_i^2 R_i^2} \tag{8.23}$$

The total mean-squared voltage on the capacitor C_i, $\overline{V_{C_i}^2}$ is the integral of $S_{C_i}(\nu)$ over the total noise equivalent bandwidth:

$$\overline{V_{C_i}^2} = \int_0^\infty \frac{4kTR_i}{1 + 4\pi^2\nu^2 C_i^2 R_i^2} d\nu = \frac{kT}{C_i} \tag{8.24}$$

Hence the *rms-fluctuation* in the number of charge carriers $\sqrt{\overline{\Delta N_c^2}}$ due to the thermal noise input voltage on the capacitor C_i follows from:

$$\sqrt{\overline{\Delta N_c^2}} = \frac{C_i}{q}\sqrt{\overline{V_{C_i}^2}} = \frac{C_i}{q}\sqrt{\frac{kT}{C_i}} = \frac{1}{q}\sqrt{kTC_i} \tag{8.25}$$

If we measure C_i in picofarads and take T as the room temperature (300 K), equation (8.25) becomes:

$$\sqrt{\overline{\Delta N_c^2}} \approx 400\sqrt{C_i} \qquad \text{(C in picofarad)} \tag{8.26}$$

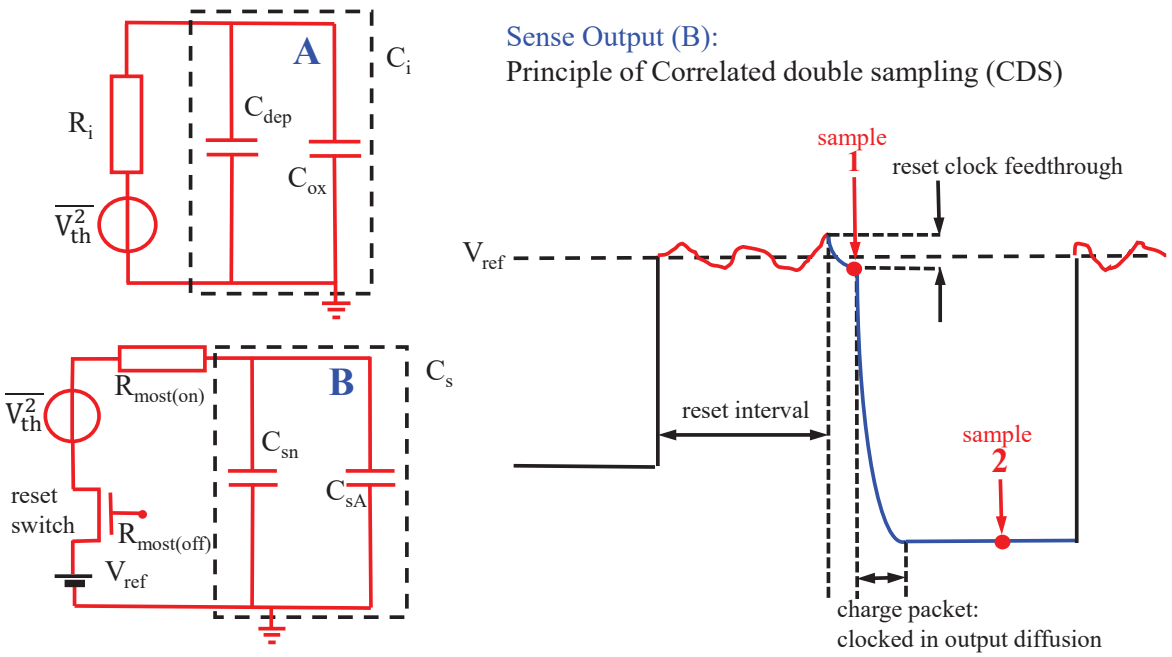


Figure 8.20: *Noise equivalent circuits for CCD input noise (A) and for sense/read noise (B). The correlated double sample principle for the latter is also shown.*

Note that this result indicates that the noise is independent of the charging resistance. For CCD's that are typical in signal processing applications the input capacitance C_i is of the order 1 pF, so this yields a value for the input noise of about 400 electrons at room temperature.
Obviously if no electrical charge injection takes place but charge carriers are photo-electrically generated in the case of an imaging device, this thermally generated so-called kTC-noise is not present.

Transfer noise

There are two principal origins of transfer noise:
- Noise due to imperfect transfer efficiency of the number of charges (also called transfer loss)
- Noise due to random capture of charge by fast interface states.

Transfer noise is caused by the fact that there are random fluctuations in the number of charges transferred between one well and the succeeding well. If a well contains N_c charge carriers, then *on average* ϵN_c will be left behind at each transfer (charge transfer inefficiency), but this will be subject to shot noise (Poisson process) with a variance ϵN_c. Since this shot noise is associated with the charge packet both when it arrives and when it leaves the potential well, the total variance associated with a well transfer amounts to $2\epsilon N_c$. If the total number of transfers equals n and ϵ is constant the total rms-fluctuation in the transfer process amounts to:

$$\sqrt{\overline{\Delta N_c^2}} = \sqrt{2\epsilon n N_c} \tag{8.27}$$

If we consider for example a two-phase 100-bit SCCD with an ϵ-value of 10^{-4} with a charge carrying capability N_c of 10^4 to 10^6 electrons, $\sqrt{\overline{\Delta N_c^2}}$ for this device would amount to 20 to 200 electrons.
The noise associated with fast interface states arises from the fluctuations in the total number of carriers that are trapped at any one moment. The mean square fluctuation of the total trapped charge per unit area $\overline{\Delta N_t^2}$ can be shown to equal approximately $0.7kTN_{ss}$ with N_{ss} the number of surface states $\mathrm{cm}^{-2}\mathrm{eV}^{-1}$. Like in the case of the charge transfer inefficiency (see above), also here the fluctuations in the number of charge carriers occur both on transfer into and out of the potential well, hence after n transfers the rms-fluctuation in the number of charge carriers due to trapping can be expressed as:

$$\sqrt{\overline{\Delta N_t^2}} = \sqrt{1.4kTnN_{ss}A_G} \qquad \text{with } A_G \text{ the area of each CCD gate electrode} \quad (8.28)$$

Typically $\sqrt{\overline{\Delta N_t^2}}$ amounts to 1000 to 5000 electrons for a 256-bit SCCD. Such *surface state* trapping does not occur in BCCS's. In bulk channel structures there is however an equivalent trapping phenomena by *bulk states*, but owing to the much lower densities of such states the trapping noise in BCCD's is typically at least an order of magnitude less than in SCCD's.

Sensing/read noise

The read-out process usually involves a reset of the output diffusion to a reference potential once every clock period. The read or sensing noise is in that case dominated by the resetting procedure, for the read-out terminals see figure (8.10). The noise equivalent circuit, figure (8.20), for the removing of charge (i.e. reading the signal) from the sense node is effectively a capacitor being charged via a resistor: the resistive channel of the reset MOST. Consequently the familiar kTC expression for the noise variance applies with C_s the total sense node capacitance, i.e. the sense node capacitance in parallel with the gate capacitance of the sense amplifier. With careful design C_s could be made $\leq$ 0.1 pF and hence we have $\sqrt{\overline{\Delta N_c^2}} \approx 120$ electrons. A technique that is frequently applied to greatly reduce this level of the reset noise is designated *correlated double sampling*.

Correlated Double Sampling (CDS) is a noise reduction technique in which the reference voltage of a pixel (i.e. after it is reset) is subtracted from the signal voltage of the pixel at the end of each integration period, to cancel the reset induced kTC noise. Figure (8.20) shows this reset MOST resistance and the associated sense node capacitance C_s. The time constant of resetting the output capacitance ($R_{most(\mathbf{on})} \cdot C_s$) is typically $10^4\Omega \cdot 10^{-13}\mathrm{pF} = 10^{-9}$ sec. So the system will have reached equilibrium in typically a few nanoseconds after the reset MOST has been turned on. As the reset MOST is turned off, there will occur a drop in the output voltage due to reset clock feed through, there will also be a much longer settling time because of the much increased time constant $R_{most(\mathbf{off})} \cdot C_s$, typically $10^{12}\Omega \cdot 10^{-13}\mathrm{pF} = 0.1$ sec. The rest of the read-out sequence is now for the pixel charge packet to be clocked into the output stage. This

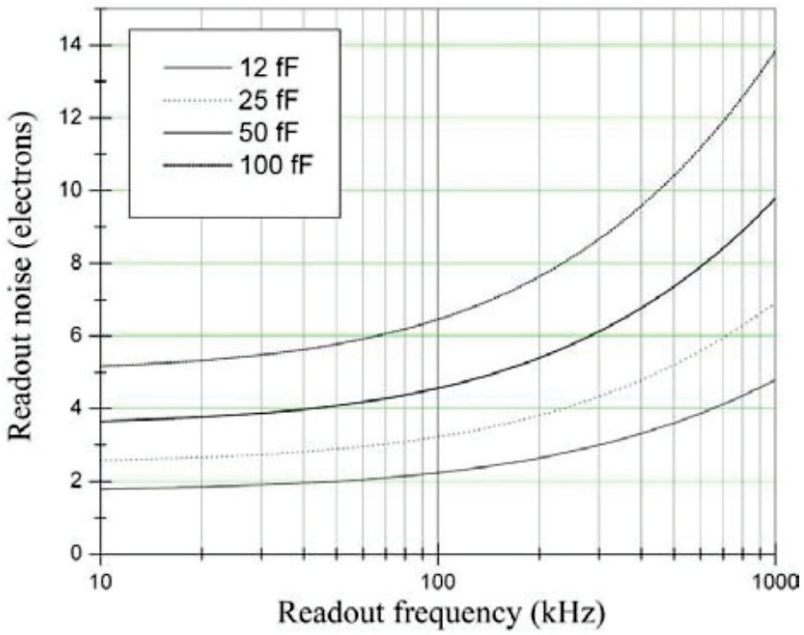


Figure 8.21: *Read noise (number of electrons) resulting from the application of correlated double sampling (CDS) as a function of output capacitance (femtoFarad) and read-out frequency (kHz).*

entails a reduced bias voltage sampled at instant **2**, see figure (8.20), relative to a fixed reference. The reference in this case is the noise voltage $\sqrt{kT/C_s}$ on top of the reference voltage established during reset. So this noise component is effectively present in the voltage level at **2**. However if two voltage level samples are taken, one at instant **1** just after the reset MOST is turned off and the other at instant **2** after the signal charge has been clocked in the output stage and assuming that the time interval between **1** and **2** is $\ll$ than $R_{most(\mathbf{off})} \cdot C_s$, then the same noise will be present in both samples. In other words, the noise is correlated and can be removed by subtracting $V(\mathbf{1})$ from $V(\mathbf{2})$, this process is called *correlated double sampling*. The subtraction process is accomplished by inverting and storing the reset voltage sampled at **1** for subsequent comparison with the signal voltage sampled at **2**.
Providing this technique in practice is fairly involved and may necessitate lowering of the CCD operating frequency, it is therefore only applied when very low noise operation is required, like single photon detection. Figure (8.21) shows an example of reduced read noise as a function of sense capacitance and read-out frequency for a CCD equipped with CDS read-out circuitry. It is clear from this picture that a read noise level of a few electrons is attainable.

Dark Noise, Fixed Pattern Noise

Dark current arises from the thermal generation of charge carriers in the CCD semiconductor (silicon) layer and can, in first instance, be considered to have two components. Firstly, the thermal generation is a random process throughout the semiconductor which gives rise to spatially white noise, i.e. the dark current induced charge packet of average

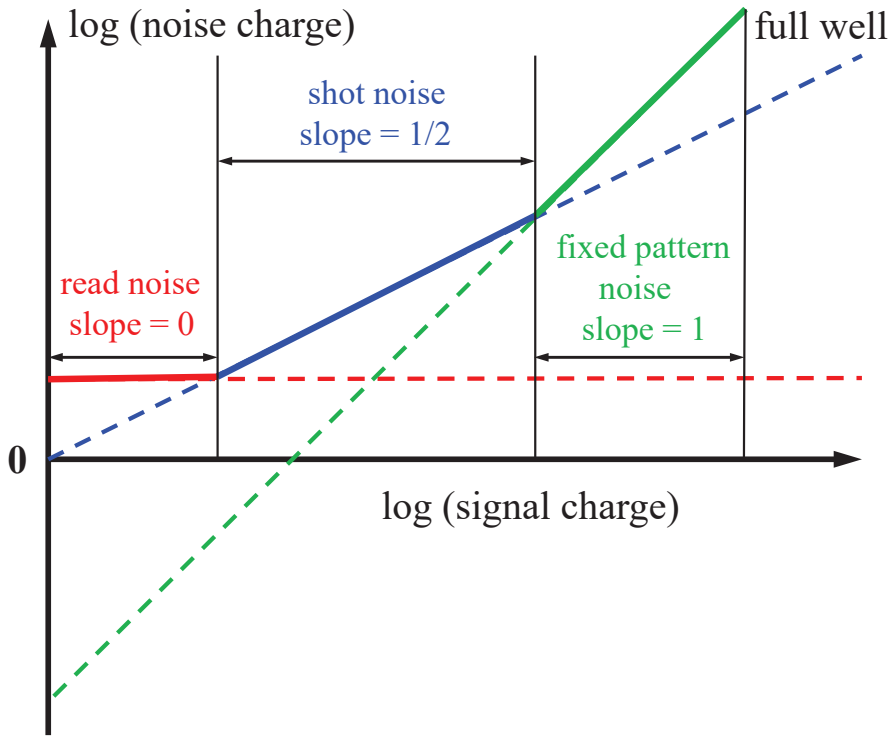


Figure 8.22: *Dominant noise charge in a CCD as a function of the signal charge stored in a potential well. The read noise is independent of signal charge, the shot noise curve has a slope of* 1/2 *on a log-log scale, signifying the square root dependence on charge carrier statistics, the fixed pattern noise (FPN) becomes potentially most prominent for large signal values nearing full well capacity. In the FPN case the noise curve has a slope of* 1*, i.e. noise charge proportional to signal charge = 45-degree line on a log-log scale. The total rms-noise charge follows from:* $Q_{TOT_n} = \sqrt{Q^2_{read_n} + Q^2_{shot_n} + Q^2_{fpn_n}}$.

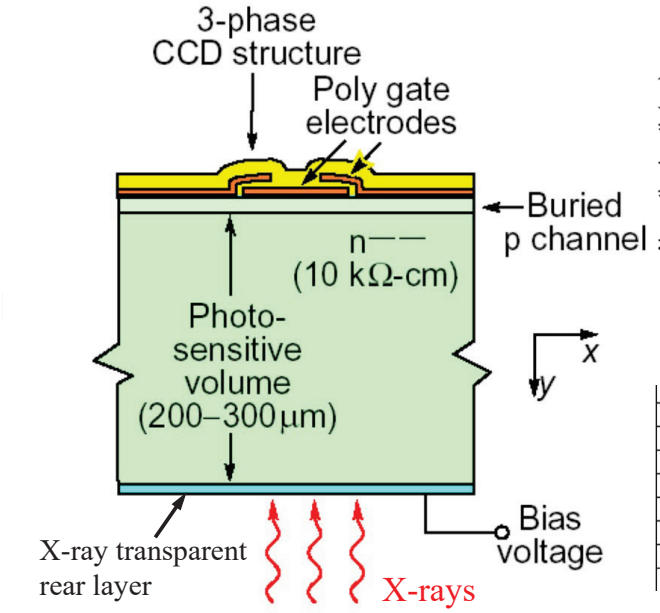
size $\overline{N}_{dark}$ gives rise to a variance $\overline{N}_{dark}$. This variance is marked as dark noise since it arises without any exposure of the CCD to radiation. A typical dark current for a good quality device at room temperature amounts to 10^{-8} A cm^{-2}, so for a typical gate transfer electrode size of 10x20 μm a potential well would collect in 100 milliseconds on average about $2 \cdot 10^4$ electrons. This 'dark' charge packet introduces a *dark noise* standard deviation of $\sqrt{2.10^4} \approx 140$ electrons.
Secondly and a potentially more seriously, this generation may not be uniform across the device. Localized regions may exhibit an enhanced generation rate of dark current introducing an inhomogeneous dark noise distribution over the device that will manifest itself prominently as fixed pattern noise specifically after long integration times. Generally such dark current spikes are associated with defects in the silicon crystal structure. The number of defects present in the processed silicon slice is a function of the quality of the starting material and the high temperature processing steps to which the slice is subjected.
As discussed in chapter 3, the generation of dark current charge carriers is a strong function of the chip's operational temperature and, consequently, its effect can be dra-

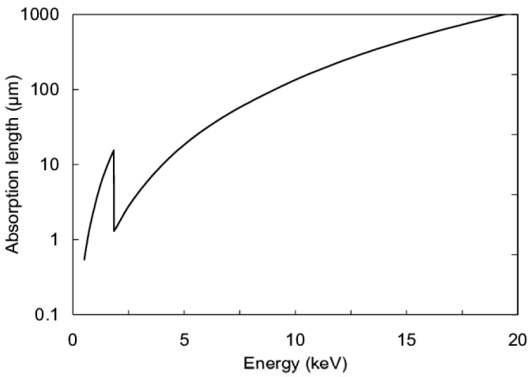
matically reduced by cooling. As derived in chapter 3, the dependence of the number of dark current charge carriers on temperature is $\propto T^{3/2}e^{-E_g/(2kT)}$ with E_g the silicon gap energy, the ensuing *dark noise current doubling temperature interval* amounts to approximately *6 degrees*. In most applications cooling down to -100 to -150 degrees Celsius will render the dark noise negligible (a few electrons) as compared to the shot noise.
Another form of fixed pattern noise will arise from slight inhomogeneities in the physical size of the pixels, this will become apparent when employing large potential well capacity since the effect is linearly proportional to the signal charge.
The collective contribution and importance of the main CCD noise sources for image application, sense/read noise, shot noise and fixed pattern noise, as a function of signal charge is shown in figure (8.22).

8.1.9 CCD as imaging spectrometer, X-ray single photon detection

Application of CCDs in X-ray imaging has undergone impressive developments over recent decades, in particular for high-energy astronomy where the X-ray photon flux is sufficiently low to separately register the (small) charge packet associated with a single X-ray photon. In contrast to the optical application where thousands of optical photons are collected in a single pixel, an exposures to X-rays should yield no more than one X-ray photon per (a few) 100 pixels, since in that case both spectral and spatial information can be obtained simultaneously. The magnitude of the charge packet represents the energy of the single incident X-ray photon. Releasing a single electron hole pair in silicon via a direct transition requires on average an energy of ≈ 3.65 eV, hence detection of X-ray photons in the energy range 200 eV to 10 keV will effectively release, on average, between 55 and 2750 charge carriers. Since the full well capacity amounts to $\approx 10^6$ charge carriers, the charge signal of a single X-ray photon will be quite small as compared to the full well capacity. Therefore it is mandatory for an accurate energy measurement that the noise level, i.e. primarily dark noise, read noise and shot noise associated with the single-photon-generated charge cloud is at the lowest possible level.
Standard CCDs have depletion depths of between 3 μm and 10 μm, and for X-ray applications special "deep depletion" devices are constructed on high-purity silicon that can provide depletion depths of 30 μm to 300 μm, depending upon silicon type and CCD design and bias. Figure (8.23)shows the average absorption depth in silicon as a function of X-ray energy. Also shown is the layered built-up for a back-illuminated fully depleted buried channel X-ray CCD. Notice in the table with performance parameters at the lower right a read noise level of 5 electrons achieved by applying CDS and an almost negligible level of dark current even for relatively long exposures. Employing back-illumination avoids the problem of having to penetrate the gate structure and the oxide layer, the shown dead layer constituting the X-ray transparent rear window can be made as thin as 50 nanometers, yielding superior response to low-energy X-rays and excellent long X-ray wavelength sensitivity. This makes these devices ideally suited as an imaging spectrometer behind a grazing incidence X-ray telescope.

Moreover, as importantly, the fully depleted Si-layer minimizes the effect of charge





Item	Measured Value
Readout Noise	4.6 e^-
Dark Current	0.002 e^-/pixel/sec
Column Defects	0
White Spots (> 100 e^-/hr)	72
Black Spots (< 80 % mean)	279
Parallel CTE	>0.999995
Serial CTE	>0.999995
Full Well	>250 Ke^-

Figure 8.23: *(top right) Average absorption length in μm for X-rays in silicon as a function of photon energy. (left) Layered BCCD structure for X-ray back-illumination. The X-ray transparent rear window can be made as thin as 50 nanometers, providing excellent low-energy X-ray response, moreover both p-type and n-type silicon layers are fully depleted. The high resistivity n-type material (10 kΩ·cm) accommodates deep depletion towards a few hundred μm's for the bulk material. Performance parameters for a 2048 x 4096 imager with a 15 μm pixel pitch at a temperature of 150 K are summarized in the table at the lower right.*

diffusion of the X-ray-generated charge cloud, since the electric field present in the deep depletion layer causes this cloud to quickly drift into the potential well. In this way degradation of spatial and spectral resolution is minimized. Another potential cause of degradation of spatial and spectral resolution is charge loss when X-ray photons are absorbed at the boundary of two or more pixels: in that case charge splitting between these pixels will occur, so-called *split events*. However this problem can be resolved later on during image processing, provided no charge loss occurs due to recombination (e.g. near a highly-doped stop-diffusion). With increasing X-ray intensity, the probability of 'overlap' between adjacent events strongly increases to the detriment of the energy resolving power of the device, hence a surface density of X-ray photons $\leq$ than one photon/(few)100 pixels should be vigorously pursued to avoid any degradation of the intrinsic spectral resolving power.
The energy resolution of a CCD is determined by the following limiting factors:

- The fundamental statistical fluctuation in the number of produced charge carriers, the so-called Fano limit. Consider an X-ray photon with energy E_x, we then have

for the average number of charge carriers $\overline{N}_c$ and its fluctuation σ_F:

$$\overline{N}_c = \frac{E_x}{3.65} \quad \text{charge carriers,} \quad \sigma_F = \sqrt{F \cdot \overline{N}_c},\ F = \text{Fano factor} \approx 0.12 \qquad (8.29)$$

If $\overline{N}_c$ is large, the probability distribution function is Gaussian with a Full Width at Half Maximum (FWHM) $\Delta E = 2.355\,\sigma_F = 2.355\sqrt{F \cdot E_x/3.65}$. This implies an energy resolving power $R = E/\Delta E \approx 9$ at 200 eV and $R \approx 64$ at 10 keV.

- Noise associated with the read-out of the CCD, which is governed by the capacitive loading of the sense output, see the discussion on read/sense noise in CCD's earlier. By using correlated double sampling this noise component can be reduced to a level of a few electrons, see figure(8.21). For a good quality CCD, the Fano shot noise and the read-out noise are the most important contributors to the energy resolution:

$$\sigma_E = \sqrt{\sigma^2_{read} + \sigma^2_{Fano}} \quad \text{often expressed in number of electrons} \qquad (8.30)$$

- Dark current and the associated dark noise. As we discussed earlier, this can be suppressed dramatically by cooling and can normally be reduced to a negligible level for integration times up to seconds. However dark current can still constitute a problem in scientific applications of CCD's, like the single photon mode for X-ray imaging spectrometry. We are dealing here in case of astronomical applications with very long integration times (several minutes to hours) and relatively slow read-out frequencies due to signal processing like correlated double sampling (read-out speed typically 50 kHz). Two solutions:
-deep cooling down to -100 $\Rightarrow$ -150 ^{0}C
-suppression by pinning (inversion)
A new technology which significantly reduces the dark current generation rate and, in turn, relaxes cooling requirements for the CCD to the level where thermoelectric cooling can be used in most applications. The technique is called Multi Pinned Phase (MPP) CCD technology.
In CCD imagers there are three main sites of dark current generation. These are : thermal generation and diffusion in the neutral bulk, thermal generation in the depletion region and thermal generation due to surface states at the silicon-silicon dioxide interface. Of these regions, the contribution from surface states is the dominant contributor for multi-phase CCDs. Dark current generation at this interface depends on two factors, namely the density of interface states and the density of free carriers (holes and electrons) that populate the interface. Electrons that thermally "hop" from the valence band to an interface state (sometimes referred to as "a mid-band state") and to the conduction band produce a dark e-h pair. The presence of free carriers will fill interface states and, if the states are completely populated, will suppress hopping and conduction and substantially reduce dark current to the bulk dark level. Normal CCD operation depletes the signal channel and the interface of free carriers, maximizing dark current generation. Under depleted conditions, dark current is determined by the quality of

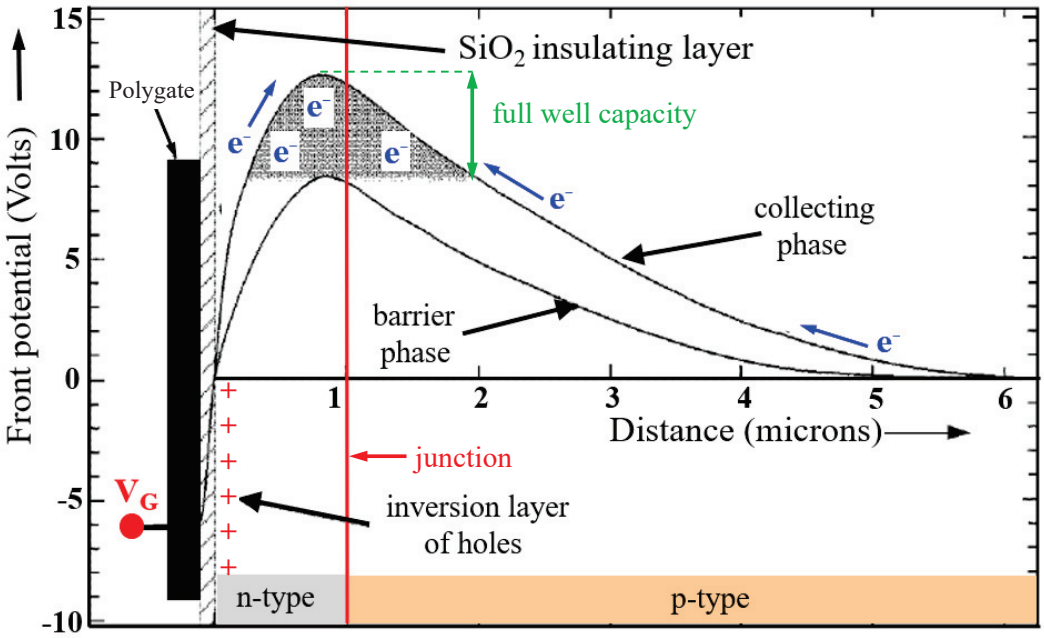


Figure 8.24: *MPP Operating Principle. A schematic showing the ideal (electric) potential profile through a front side illuminated CCD whose front surface is inverted with multi-pinned phase (MPP). The CCD consists of a polysilicon gate, which forms part of the electrode structure, a surface layer of oxidized silicon (SiO_2) and the epitaxial layer which comprises p-doped silicon with an n-doped buried-channel for charge transfer. MPP pins the surface potential by populating the Si-SiO_2 interface with holes. The holes passivate the Si-SiO_2 interface states and create an electric field which directs signal charge away from the interface towards the buried n-channel.*

the silicon-silicon dioxide interface or the density of mid-band states. The dark current can be suppressed by means of so-called pinning (=inversion). The operating principle of multi-phase-pinning (MPP) is shown in figure (8.24). Consider a three-phase n-buried channel CCD with phases $\phi 1, \phi 2, \phi 3$. If we apply a gate voltage $\Phi_G \ll 0$ to $\phi 1$ and $\phi 3$, a deep potential minimum for holes will form at the Si-SiO_2 interface that will be filled with holes from the p^+ channel stop diffusions. This creation of an inversion layer of holes (minority charge carrier in n-type silicon, see also the relevant paragraph in the treatment of the MIS element), causes thermally generated 'dark' electrons in their surface layer to immediately recombine and, hence, they will not contribute anymore to the surface layer dark current. One could say that the potential of the Si-SiO_2 interface has effectively been 'pinned' to the substrate potential. During integration, it is desirable to bias all the gates into inversion. However, in a normal CCD with all gates inverted, there is no barrier to separate pixel changes, so this technique can only be applied to two out of three electrodes and will yield a dark current reduction of a factor three.

In the MPP device, during manufacturing an extra ion implantation step is incorporated under one of the electrodes (usually $\phi 3$). This implant creates a built-in

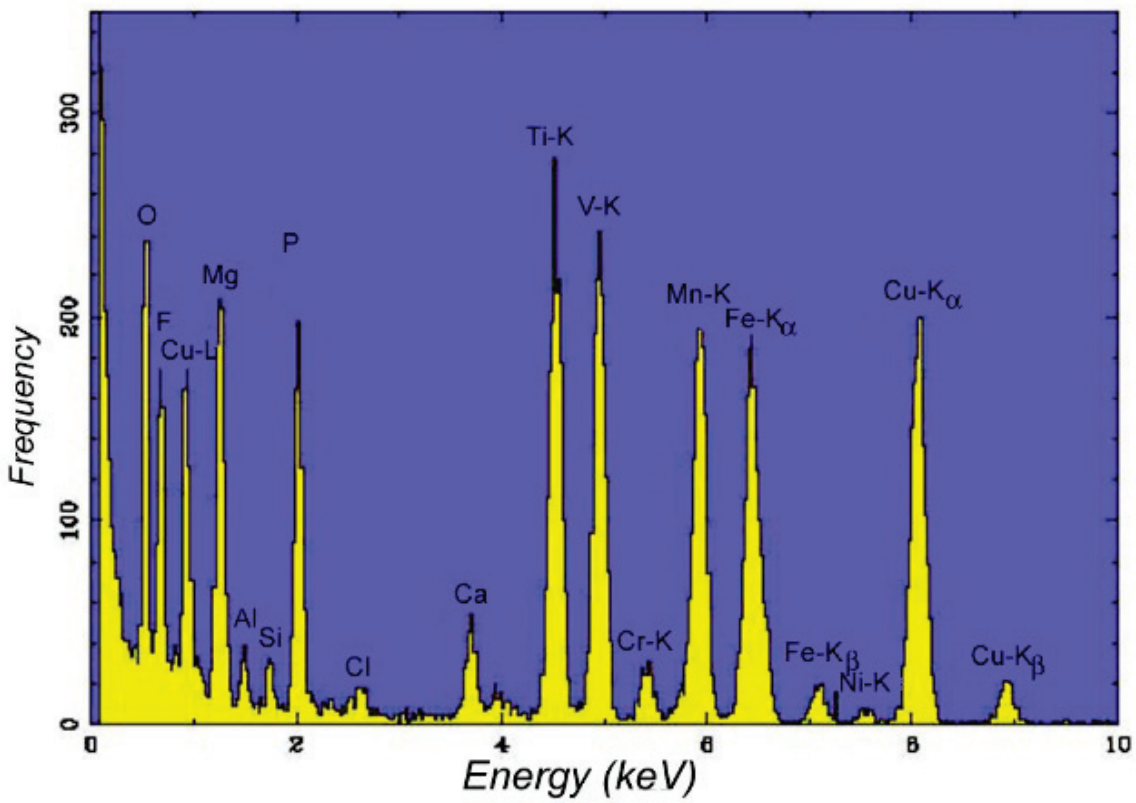


Figure 8.25: *X-ray spectrum from a variety of elemental fluorescence lines between Oxygen at 0.5 keV and Copper at 8 and 9 keV, demonstrating the energy resolving power of an X-ray photon counting CCD.*

potential barrier in each pixel, which allows charge integration to be performed with all of the vertical clocks biased into inversion. Under these conditions, charge collects under $\phi 1$ and $\phi 2$ while remaining separated from neighboring pixels. The only disadvantage of the implant MPP technique is that it substantially reduces the full well capacity per pixel as compared to a normal CCD, but this is of no significance for the X-ray spectrometric application since only a minor fraction of the well capacity is used by a single photon charge cloud. MPP-CCD technology has achieved dark floors of 10-50 pA cm^{-2}, a factor of 200-1000 times lower as compared to non-pinned CCD's.

- The efficiency of charge transport, the Charge Transport Efficiency (CTE), is a key parameter for retaining the (Fano) shot noise dominated energy resolving power. We have CTI = 1-CTE, with CTI the charge transport inefficiency, that is defined as the relative charge loss per individual pixel shift. This charge loss is again caused by charge traps like lattice defects and contaminants that can capture charge, at least temporarily. X-ray CCD's require a very high CTE value (CTI = 10^{-5} – 10^{-6}) to retain the intrinsic, quantum noise dominated spectral resolution. The CTE can deteriorate as a result of lattice defects that are caused by ionizing

radiation, in particular protons. This is particularly relevant for operation of the CCD in outer space in case it repeatedly passes through the radiation belts and under influence of energetic solar flares. This radiation damage can, at least partly, be recovered by an annealing process at a temperature of 130 ^{0}C. Protection against radiation damage requires embedding of the image sensor inside proper (Z-value) shielding material and limitation of the material volume through which the charge transport is taking place.
Figure (8.25) shows a CCD registered X-ray spectrum obtained in single photon counting mode. It shows the energy resolving power for a host of elemental fluorescence lines from Oxygen (O) near 0.5 keV all the way up to the Iron (Fe), Nickel (Ni) and Copper (Cu) complex between 6 and 9 keV.

- High quantum efficiency at low X-ray energy: Back-illumination
In order to raise the detection efficiency for soft X-rays substantial traversion of the front gate structure and the oxide layer needs to be avoided and the CCD needs to be illuminated from the rear side by etching away the p-substrate nearly up to the boundary of the depletion layer in the p-type semiconductor, a process called 'thinning'. The high resistivity p-material (allowing for large depletion depths) has a typical acceptor concentration of $N_a \approx 10^{19}$ m^{-3}. After the etching process, because of natural oxidation of the rear side. a potential dip for electrons will develop in the vicinity of the backside of the waver. Moreover the oxidized backside does no longer have a properly defined potential, owing to a myriad of surface traps and an inhomogeneous crystal surface of the oxide layer. Therefore a thin p^+ layer with high conductivity needs to be implanted at the backside of the device with an acceptor concentration roughly 1 million times higher than the depleted p-layer, i.e. $N_a \approx 10^{25}$ m^{-3}. The p^+ implant should have a thickness of 25-50 nanometer to accommodate a transmission of ≥ 80 % above 0.3 keV X-ray photons. This p^+ backlayer provides a very thin field-free boundary layer and introduces the following benefits/drawbacks:
- Its high conductivity secures a well defined potential at the backside of the device.
- The difference in doping concentration between the p- and p^+-layer creates a potential barrier between the p-depletion layer and the backside that prevents charge loss by diffusion towards the backside.
- The p^+-layer is not really an inactive dead layer: its quantum efficiency varies from 0 at the back surface to 1 at the p^+/p boundary. In case an X-ray photon gets absorbed in the p^+-layer, only part of the generated charge cloud will be detected, this leads to so-called *partial events*. They will show up as a flat plateau at the low energy side of the full absorption photo-peak in the X-ray spectrum. From the measured fraction of partial events relative to the photo-peak events one can deduce the thickness of the p^+-layer. Figure (8.26) shows the various potentials near the backside before and after implantation.
To form the p^+ passivation layer in a back-illuminated CCD, an ion implantation process called a 'laser doped' anneal is utilized, which involves the following steps:
1. Boron atoms are incident onto the back surface of the CCD to implant them into the silicon. The energy of the atoms determines the average depth that the

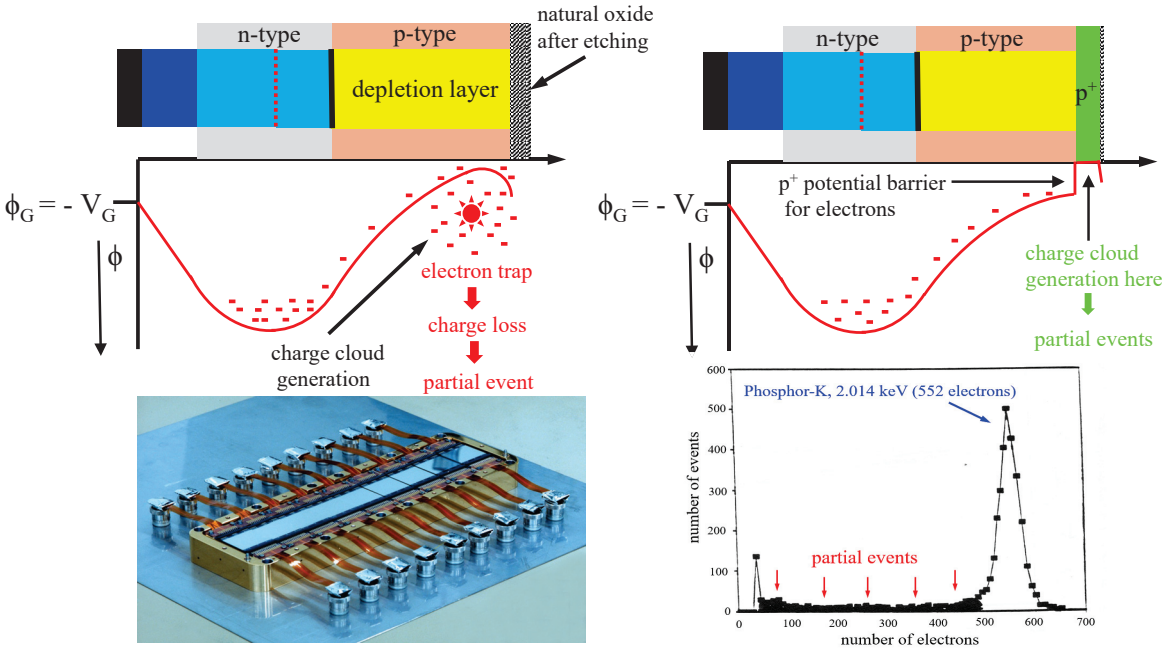


Figure 8.26: *Potential distributions in backilluminated CCD's. (Upper left) After thinning to nearly the depletion layer boundary in the high resistivity ($N_a \approx 10^{19}$ m^{-3}) p-material. The natural oxide layer that causes a potential dip for electrons at the backside is indicated. (Upper right) After a high doping implantation with Boron a p^+-layer is created with $N_a \approx 10^{25}$ m^{-3}. (Lower left) A strip of Boron implanted (GILD) backilluminated CCD's used for the read-out of the X-ray spectra obtained with the XMM-Newton X-ray Space Observatory. In addition the CCD's are covered with a thin Al-window, varying in thickness between 45 and 75 nm, to block incident UV-radiation. (Lower right) Partial events originating from photon absorption in the implanted p^+ layer.*

atoms will penetrate into the silicon which is the dominant contribution to the thickness of the p^+ layer. The process to passivate the back-surface implants boron at energies between 10 and 25 keV.

2. The back surface of the device is rapidly melted using a UV laser pulse (wavelength 248 nm, 30 ns pulses) to provide a shallow anneal. Boron has a high diffusivity in liquid silicon and as the silicon cools and re-crystallizes the boron atoms are incorporated into the silicon lattice forming the passivation layer.

In any ion implantation process there is an element of lattice damage associated with the implantation and this can create non-uniformities in the back of the device and create additional generation/recombination centres and defects in the silicon lattice.

An alternative technique is to form the p^+ layer using Gas Immersion Laser Doping (GILD). GILD passivates the surface of a back-illuminated CCD with a thin boron implant (25 to 50 nm from the surface) in order to improve the soft X-ray response of the detectors. GILD is implemented through the following process:

I. The CCD is held in an atmosphere consisting of diborane gas (B_2H_6) and a pulsed KrF excimer laser (wavelength 248 nm) is stepped over the CCD in order

to dissociate the diborane gas and melt a very small layer into the silicon.
II. The atomic boron penetrates into the molten silicon by liquid phase diffusion and the dopant is introduced as far as the melt depth.
III. The profile of the boron doping is determined by the laser energy.
The GILD process is able to produce back-illuminated CCDs with p^+ layer of $\approx$ 25-50 nm, but since the CCD has to be in a diborane atmosphere, it takes longer to process multiple CCDs compared to ion implantation techniques, but there is no un-annealed lattice damage created as the boron atoms do not require activation. The GILD process is only possible on a single device at a time and so it has now been superseded by the ion implantation technique.
There are several other passivation techniques, apart from the ion implantation process described here, that have been successfully developed over recent years, however we shall not further elaborate on these here.

8.2 CMOS (Complementary Metal Oxide Semiconductor) Imager

8.2.1 Main differences with CCD's

CMOS image sensors are, in contrast to CCD's, fabricated in "standard" CMOS technologies. Their main advantage over CCDs is the ability to integrate analog and digital circuits with the sensor on pixel level. This results in a number of advantages over CCD's:
- Less chips used in the imaging system
- Lower power dissipation
- Faster readout speeds
- More programmability
- New functionalities like high dynamic range, biometric applications etc.

However CMOS imagers also have distinct disadvantages as compared to CCD's:
- In general they have lower image quality performance than CCD's
- Standard CMOS technologies are not optimized for imaging applications
- More circuits do result in more noise, in particular Fixed Pattern Noise (FPN)

8.2.2 Basic pixel architectures

We shall now briefly discuss a few aspects concerning the basic elements of CMOS imaging architectures. Three basic pixel architectures can actually be distinguished in order of increasing complexity:

- Passive pixel architecture (PPS):
  A passive-pixel sensor consists of passive pixels which are read out without amplification, with each pixel consisting of a photodiode and a MOSFET switch, i.e. one transistor for each individual pixel. So in a PPS photodiode array pixels contain a pn-junction, an inherent, integrated capacitance C_{pd} and a MOSFET

as selection transistor.
Originally, passive-pixel sensors were being investigated as a solid-state alternative to vacuum-tube imaging devices. The MOS passive-pixel sensor used just a simple switch in the pixel to read out the photodiode-integrated-charge built up during the light exposure interval. Pixels were arrayed in a two-dimensional structure, with an access enable wire shared by pixels in the same row, and an output wire shared by column. At the end of each column was a transistor. Passive-pixel sensors constitute small pixels with a large fill factor, but also suffer from serious limitations, such as low signal to noise ratio, slow readout, and lack of scalability. The photodiode readout bus capacitance resulted in increased read-noise level and suppression by Correlated Double Sampling (CDS) could not be used with a photodiode array without implementing an external memory.

- Active-pixel architecture (APS)
  The active-pixel sensor consists of pixels, each containing one or more MOSFET

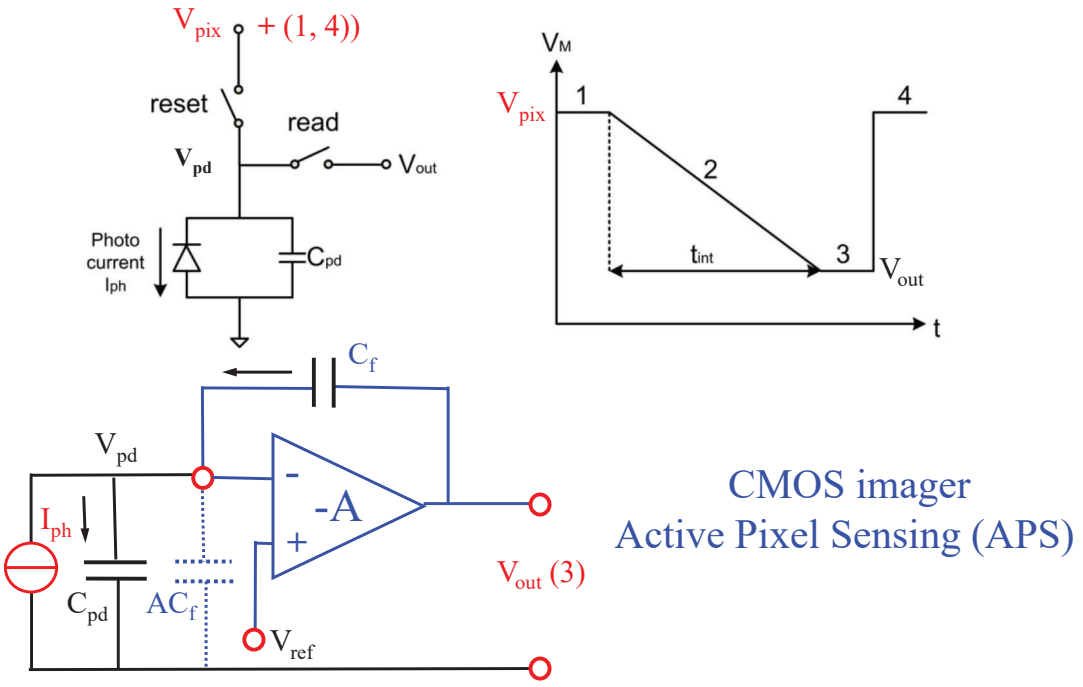


Figure 8.27: *CMOS Active Pixel Sensing. Enabling the reset switch applies a reverse bias potential V_{pix} to the photodiode pixel. After breaking the connection, the potential V_{pd} on the diode capacitance C_{pd} starts to drop due to discharge by the photocurrent which is generated by the light incident on the pixel. After a preset integration period ΔT_{int} the read switch is enabled and the remaining value of the potential V_{pd} is read out ($\Rightarrow V_{out}$) by the pixel based Capacitive Trans-Impedance Amplifier (CTIA), followed by a reset to the original V_{pix} potential. The feedback capacitor C_f can also be regarded as an input capacitance $\bar{A}C_f$ (indicated in dotted lines) without feedback, at low frequencies $\bar{A}$ is so large that the CTIA input reflects a virtual short to the photodiode pixel.*

amplifiers which convert the photo-generated charge to a voltage, amplify the signal voltage, and reduce noise. The concept of an active-pixel device was developed in 1968 with sensor arrays with active MOS readout amplifiers per pixel with in essence the modern three-transistor configuration: the buried photodiode-structure, selection transistor and MOS amplifier. Read-out of an APS architecture is much faster than for the PPS pixels with a much improved signal to noise ratio. It is currently the technology of choice, although it introduces a larger pixel with a lower fill factor than in the PPS case.
Figure (8.27) shows the principal components of an APS architecture, including the on-pixel Capacitive Trans Impedance Amplifier (CTIA). The typical photocurrent in active pixel sensors is quite small and hence current integration over a preselected time interval is implemented in operational practice. Such an operational cycle is also illustrated in figure (8.27):
- The reset switch is closed and resets the photodiode to a reverse bias potential V_{pix}, V_{pix} is the actual supply voltage for the pixel.
- Next, the reset switch opens and the reverse bias potential V_{pd} starts to drop following discharge of the photo-diode capacitance C_{pd} by the light generated photocurrent I_{pd}.
- After a preselected time time interval ΔT_{int}, a read switch closes and the remaining photo-diode potential $V_{pd} = Q_{pd}/C_{pd}$ is fed into the CTIA, converting the instantaneous value of $V_{pd} \Rightarrow V_{out}$.
In chapter 7, discussing a single photodiode, we derived the following expression for the inversion by a trans-impedance amplifier:

$$-\bar{V}_u = \bar{A}\ \bar{I}\frac{\bar{Z}_p\bar{Z}_f/(1+\bar{A})}{\bar{Z}_p + \bar{Z}_f/(1+\bar{A})} \tag{8.31}$$

Referring to figure (8.27) and replacing $\bar{Z}_p$ by $1/j\omega C_{pd}$ and $\bar{Z}_f$ by $1/j\omega C_f$ we get:

$$\begin{aligned}
\bar{V}_{out} &= -\frac{\bar{A}\ \bar{I}}{j\omega(C_{pd}+C_f)+j\omega C_f\bar{A}} \Rightarrow \bar{I} = -\frac{j\omega\bar{V}_{out}}{\bar{A}}[(C_p+C_f)+\bar{A}C_f] \Rightarrow \\
j\omega\bar{V}_{out} &= -\frac{\bar{I}}{C_f} \quad \text{if} \quad \bar{A} \gg \frac{C_{pd}}{C_f} \quad \text{with} \quad \bar{V}_{out} = |\bar{V}_{out}|e^{j\omega t} \ \Leftrightarrow\ (FT) \\
-\frac{dV_{out}(t)}{dt} &= \frac{I(t)}{C_f} \quad \Rightarrow \quad V_{out} = -\frac{\int_0^t I(t')dt'}{C_f} = -\frac{Q(t)}{C_f}
\end{aligned} \tag{8.32}$$

This signifies that the remaining photodiode charge Q_{pd} to be read-out is converted by the CTIA to a voltage source with magnitude $V_{out} = -Q_{pd}/C_f$, i.e. charge sensitive amplification.
- At the end of the cycle, the read switch first opens followed by the closure of the reset switch. That resets the reverse bias to the supply voltage V_{pix} and marks the beginning of the next frame.
Figure (8.28) shows the read-out with the aid of enabling MOSFET switches of a two-dimensional image array. Shown is an active pixel array (APS) with integration of capacitive trans-impedance amplifiers on individual pixel level.

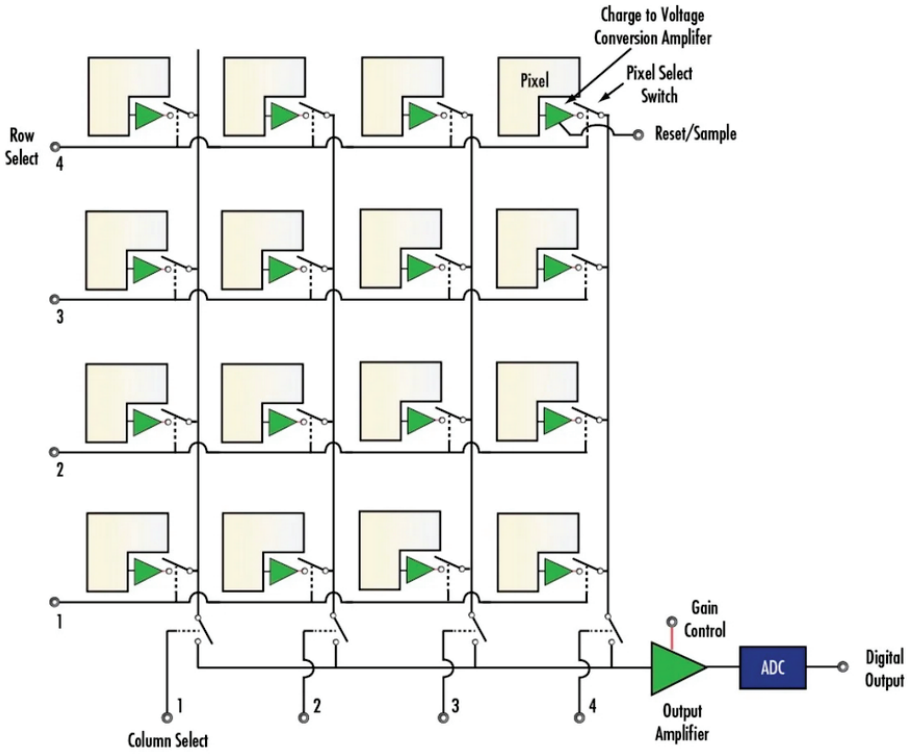


Figure 8.28: *Two-dimensional CMOS array with active pixels, the small green triangles signify the CTIA's that are integrated on pixel level. The array's of MOSFET switches indicate the selection matrix for pixel read-out on row and column level.*

- Digital pixel architecture (DPS)
  An important trend in digital imaging system design is the integration of the CMOS image sensor with analog and digital processing down to the pixel level. Such integration saves power and reduces system size and it additionally provides the ability to rethink the imaging system architecture. An important example of these architectural aspects is where and how to perform the analogue to digital (A/D) conversion. Figure (8.29) shows possible levels of integration for an Analogue to Digital Converter (ADC).
  In a Digital Pixel Sensor (DPS) each pixel (or group of pixels) has an ADC and all ADC's operate in parallel.The benefits and drawback's of such a digital pixel approach can be summarized as follows:
  - High-speed read-out due to parallel conversion and digital read-out.
  - Eliminates column temporal noise and Fixed Pattarn Noise (FPN).
  - The design process is scalable, e.g. low speed ADC's can be implemented with very few transistors.
  - Relatively large pixel size related to ADC accommodation.
  - Limited ADC resolution because of limited pixel size, this size problem becomes less severe as technology scales.
  - Considerable complexity of implementation

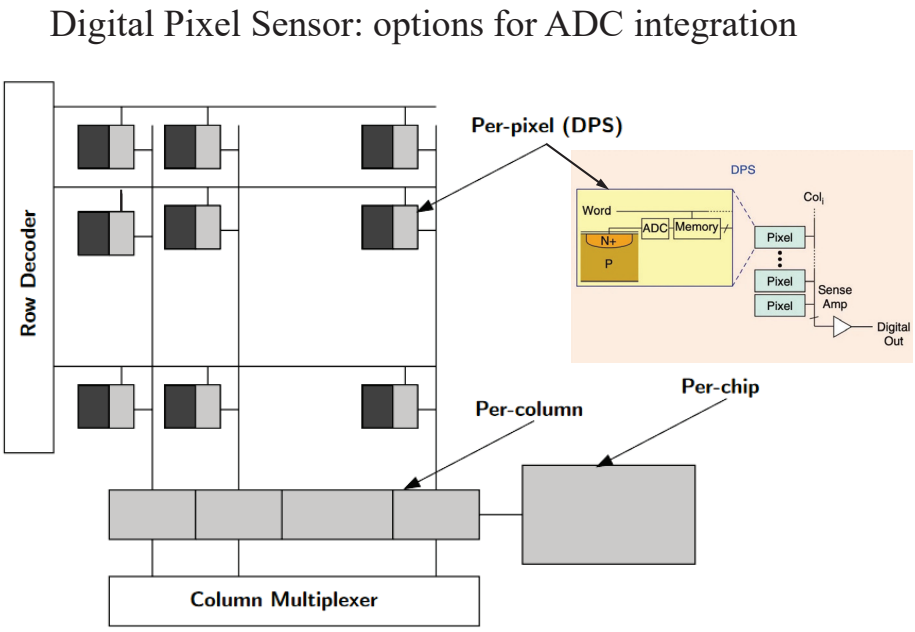


Figure 8.29: *Schematic of a Digital Pixel Sensor (DPS) and a matrix showing the options for ADC integration: pixel level, row/column level or chip level.*

8.3 Microchannelplate imagers and image intensifiers

The microchannelplate belongs, like all the other imaging sensors, to the class of modulating transducers.

8.3.1 Operational principle of a continuous channel electron multiplier

Development of continuous-dynode electron multipliers (CEM) took place in the 1960's with the production of high-resistance surfaces on lead-glass through the process of reduction in a hydrogen atmosphere at temperatures between 300 ^{o}C and 450 ^{o}C. The operation of a CEM is shown in figure (8.30), upper left. The multiplier is operated under high vacuum conditions with a high voltage established over the channel. An energetic photon or charged particle striking the wall of the channel releases an electron with a certain initial energy that is subsequently accelerated along the channel axis where, during the drift, it will strike the wall with sufficient energy to release secondary electrons. This process is repeated many times over and finally results in the generation of a charge pulse containing up to 10^8 electrons. When the voltage over the channel is increased, the energy of the electrons striking the wall will increase, but the total number of impacts will however decrease. It may therefore be expected that the charge gain (amplification) will increase to a certain maximum value and will then start to decrease. The expression for the charge gain as a function of the applied high voltage

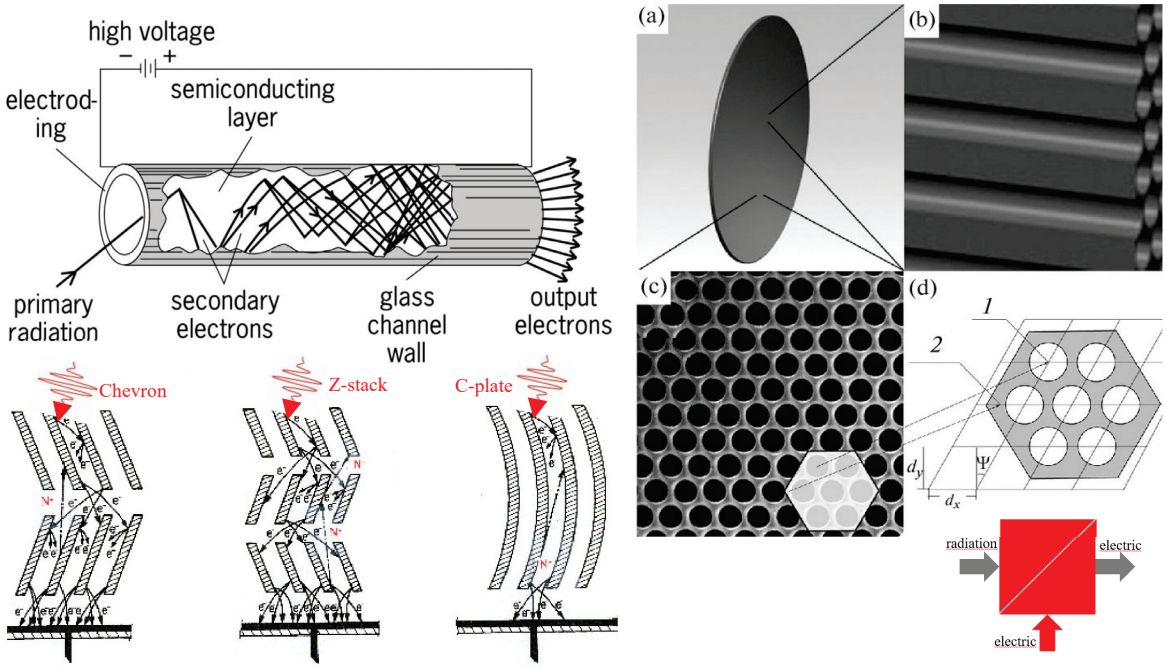


Figure 8.30: *Operational principle of a microchannelplate.*

has been derived as:

$$G = \left(\frac{C_s V_0^2}{4V\alpha^2}\right)^{4V\alpha^2/V_0} \tag{8.33}$$

where G is the gain, V_0 the energy gained by the electron traversing the applied potential difference, V is the initial energy of the secondary electron emitted normally to the channel wall, α the length-diameter ratio of the channel and $C_s = \delta/V_c$ with δ the secondary emission factor and $V_c = V_0^2/(4V\alpha^2)$ the collision energy.
However the experimental gain curve does not show a maximum, as predicted by the model, but continues to much higher gain levels and then starts to saturate. The saturation is due to the built up of an electron space charge cloud at high amplification ($\geq$ few times 10^8) and distortion of the channel electric field. The reason for this is twofold. First, contrary to the model assumption, not all secondary electrons are emitted normally to their exit plane. Second, and more important, the gain is enhanced by *ion feedback*. The applied voltage that accelerates the electrons down the channel will also accelerate a positive ion that gets created by impact of an electron that ionizes a residual gas molecule. This positive ion becomes accelerated back to the input of the channel where it can impact the wall and restart the gain process. The level of this positive ion feedback will obviously depend on the residual ambient gas pressure and on the level of the applied voltage. This feedback of positive ions can be prevented by curving the channel, forcing the ions to impact the wall already within a short distance as compared to the channel length.
A CEM with an internal channel diameter of order 1.0 mm and a length to diameter ratio of order 100:1 can produce a modal gain $> 10^8$.

8.3.2 The microchannelplate concept, stack configurations

It is important to note from equation (8.33) that the fundamental electrical characteristics of a CEM depend on the length to diameter ratio α of the channel and not on the absolute physical dimensions. The size of the channel can thus be reduced to the limit set by the available glass technology, so many channels can be bonded together to produce a detector with image recording capability. This results in a so-called *microchannelplate* that constitutes millions of glass capillaries (= channels) with diameters ranging from 5-25 μm that are melted together into a thin disk of order 1 mm thick, see figure (8.30). The inside surface of each channel involves a material with high secondary emission factor and relatively high resistance. The front and back face of the plate are coated with a thin metal film that constitutes an electrode. By applying high voltage to these thin film electrodes, each channel acts as an individual electron multiplier. The main operational properties of an MCP are summarized as follows:

- High charge amplification factor. To achieve this and to prevent positive ion feedback, the MCP's are often arranged in a stack or contain curved channels. Figure (8.30) shows the three most common configurations: the Chevron stack, the Z-stack and the curved channel or C-plate. Charge amplifications by these configurations range from 10^6 to 10^8.
- Two-dimensional image registration/amplification.
- Fast timing, better than 1 nanosecond.
- Low sensitivity for magnetic fields, in contrast to photomultiplier tubes.
- Potentially sensitive to detection of charged particles and electromagnetic radiation from infrared to X-rays by selection of the proper photocathode material, either free standing or coated on the MCP front surface.
- Capability for single photon counting.
- Low power consumption
- Physically compact and low mass, diameters standard up to 15 cm.
- Stable in dry air, with the exception of the photocathodes, the latter requires storage in vacuum or a dry Argon atmosphere.

8.3.3 Manufacturing process

The several steps for the manufacturing process of an MCP are illustrated in figure (8.31).
The starting material consists of a special type of lead glass with the following constituting elements: 50% leadoxide (PbO), 40% silicon dioxide (SiO_2) and a 10% mix of several different types of alkali-oxides. This special lead glass has intrinsically a high electrical specific resistance that can be lowered (i.e. tuned) by removing oxygen through reduction in a hydrogen atmosphere at 400 ^{o}C. Some lead will evaporate, the

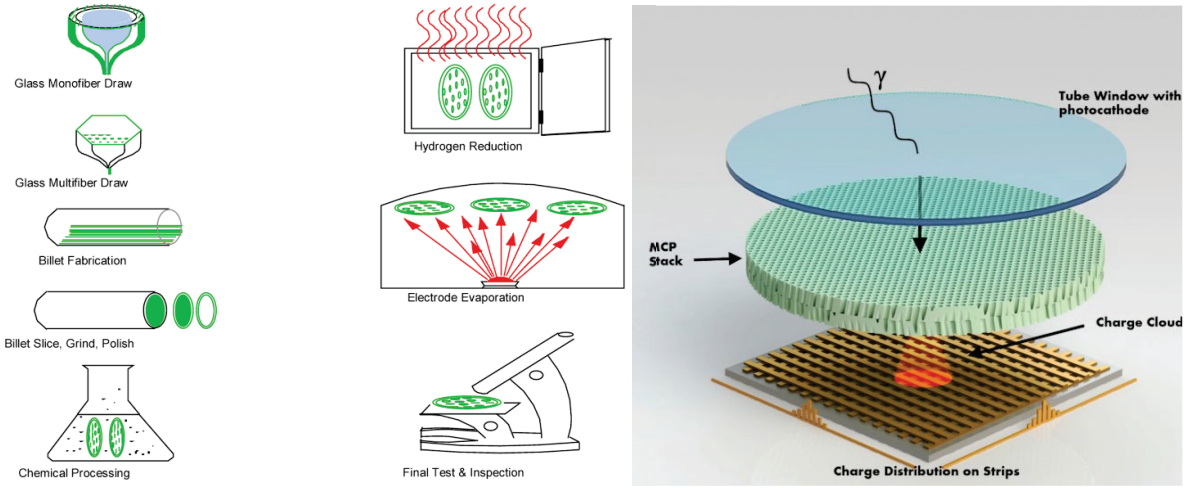


Figure 8.31: *Overview of the fabrication steps in the manufacturing process of a microchannelplate.*

remaining fraction coagulates to metal clusters. The high secondary electron emission factor is obtained within a thin surface layer containing Si, O_2, K and Pb.
The process starts with a cylindrical hollow glass tube in which a soluble glass rod is inserted. The aim of having this soluble rod is to prevent having distortion during the fiber drawing process. The soluble core and the hollow glass cylinder need to be well matched in terms of temperature coefficient of expansion, viscosity, diffusion coefficient and solubility.

- Start with a cylinder of say 40 mm diameter, draw a fiber with diameter of 0.8 mm, cut this and assemble the 0.8 mm glass fibers parts into a bundle and fuse them together at a high temperature to a hexagonal multi-fiber, see for geometry figure (8.30).
- Apply a second draw of the hexagonal multi-fiber that reduces the diameter again by a factor 40 -80, i.e. channel diameters of 10-20 μm, subsequently cut the drawn multi-fiber, and fuse the parts again together at high temperature.
- From that fused bundle, one cuts a slice at some bias angle against the fiber axes.
- The soluble fiber cores can now be etched away to obtain the pattern of holes (the channels).
- With a chemical treatment, one generates a thin semiconductor layer (see above) on the channel walls. That layer is optimized for a high efficiency of secondary electron emission and for a predetermined low level of electrical conductivity.
- One evaporates a thin metal film, typically a nickel alloy (e.g. NiCr), onto the end faces to form electrodes. Those electrodes penetrate somewhat into the channels.

The whole device must be operated in a high vacuum, e.g. at a residual pressure of 10^{-4} Pascal. Figure (8.31)(left) shows the chronological order of the various manufacturing

steps. At the right an image intensifier set up is displayed featuring a free standing photocathode, two microchannelplates in a chevron stack providing a 10^6-10^7 charge amplification and a resistive charge read-out system (see further on).

8.3.4 Charge gain and distribution function

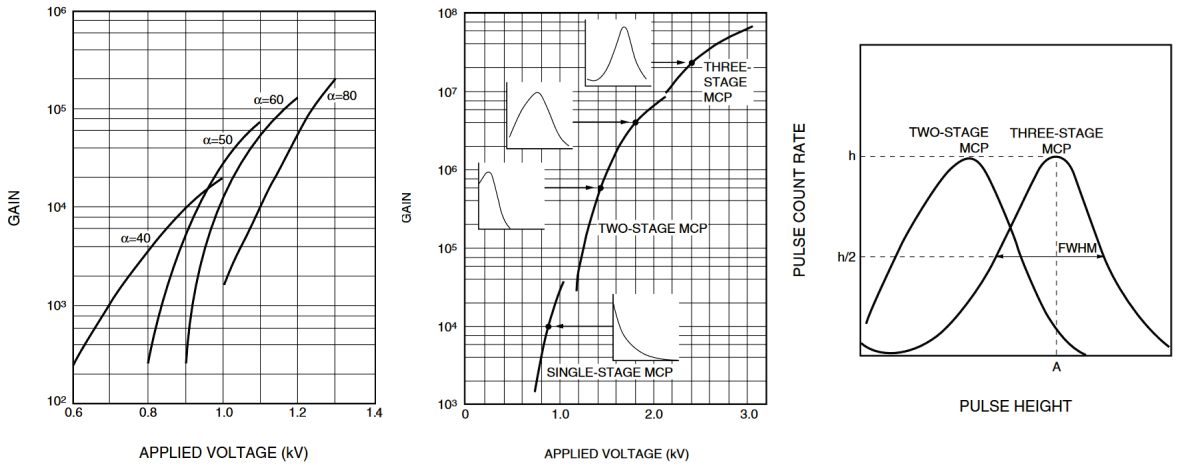


Figure 8.32: *Gain and pulse height distributions of single and stacked microchannelplates. The leftmost picture shows the gain dependence on the the length to diameter ratio α while maintaining the same channel wall characteristics (composition of the secondary emission layer). A single stage microchannelplate that is not driven into saturation shows a nearly exponential pulse height distribution (lowest in the middle picture). In the case of double and triple stage microchannelplates with high amplification, major saturation effects occur causing the pulse height distribution to shift towards a Gaussian distribution. This then allows characterization of the charge resolution by the full width at half maximum (FWHM) of the charge distribution function, see rightmost picture.*

The approximate gain of a microchannelplate can be written as:

$$G = e^{\mathbf{g}\alpha} \tag{8.34}$$

using explicitly the length-to-diameter ratio α of the channel. Furthermore $\mathbf{g}$ is the secondary emission characteristic of the channel wall and is called the gain factor. The gain factor is an inherent property of the channel wall material and is a function of the elctric field strength inside the channel. The leftmost panel in figure (8.32) shows this gain characteristic for MCP's with the same channel wall material but having different values of α ranging from 40 to 80. In general when α increases the gain reaches into the higher voltage region and increases exponentially. However if the gain exceeds approximately 10^4 the positive ion feedback becomes increasingly severe, and it is not possible to raise the gain of a straight channel single stage MCP to much larger values. Normally α-values of 40 to 60 are used, providing a gain of 10^4 at a high voltage around 1 kilo-Volt. When higher gains are required, two or three MCP's are stacked

in proximity with their bias angles relative to the channel axis alternately opposite to each other. This leads to the aforementioned Chevron and Z-plate configurations. The gains offered by single, two and three stage MCP's are displayed in the middle panel of figure (8.32), together with the associated pulse height distributions. A single phase MCP shows approximately an exponential distribution without any peak:

$$n(\mathbf{Q}) = n_0 e^{-(\mathbf{Q}/\overline{\mathbf{Q}})}, \text{ with } \overline{\mathbf{Q}} \text{ the average gain, a function of the voltage } V_{MCP}) \quad (8.35)$$

with as normal operation range V=500-1200 Volt, $\overline{\mathbf{Q}}$= 10^3-10^4.
The two stage MCP's offer a gain higher than 10^6 and the three stage MCP's a gain in excess of 10^7. In both these cases the total gain is lower than the multiplied gain of each individual MCP because charge loss occurs when charge moves through each MCP and also saturation becomes more severe when the total charge in the MCP channel drastically increases. The characteristic pulse height distributions at high gains where saturation effects start to play a significant role changes from exponential to peaked distributions and tend to a Gaussian shape. These peaked high gain distributions can be characterized by peak gain, $10^6 \Rightarrow 10^8$, and a characteristic width represented by the full width at half maximum (FWHM), see the rightmost panel of figure (8.32).
In case the inter-plate voltage between successive stages is increased, the radial spread of the charge cloud between the plates will be minimized. The number of channels that will be triggered in the plate that is next inline will be reduced, however the charge/channel will be increased and saturation will become more severe. This results in two effects: the overall gain will go down, but the charge distribution becomes narrower which improves discrimination against 'dark' shot noise originating from the plate walls and electronic noise. Figure (8.33) show high gain pulse height distributions for a Z-plate (stack of three straight channel MCP's) and a curved channel C-plate MCP's. Both show very pronounced peaked charge distributions with excellent potential for

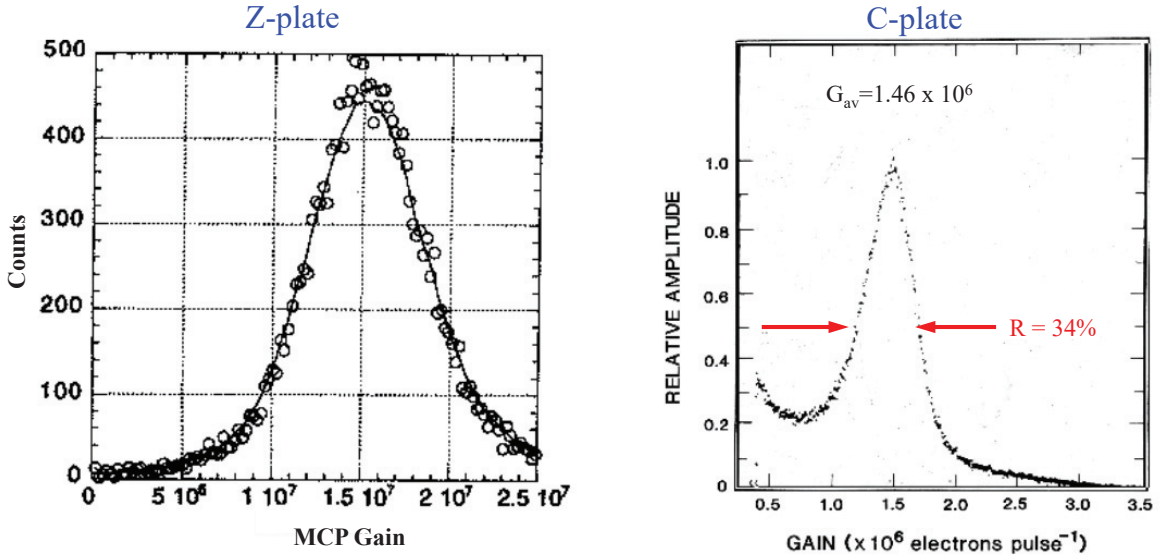


Figure 8.33: *Charge distributions at the output of a Z-plate and a single curved channel C-plate*

discrimination against noise.

8.3.5 Charge read-out systems

The charge pulses exiting the rear side of the plate need to be detected with the aid of a charge read-out system. We can distinguish between two main categories:

- Conversion of the charge pulse to an optical signal with the aid of a suitable phosphor. In this phosphor the incident electron pulse energy is transformed in optical light by the process of luminescence. The light picture created in this way can be registered by a CCD or CMOS device and be converted to an electronic image. To prevent spreading of the electron beam leaving the rear side of the microchannelplate, a positive potential is applied between the back surface of the plate and the light emitting phosphor. In the chevron stack and the Z-plate stack, some loss in image resolution will be suffered as a result of the charge spread between the individual plates that are nominally at a mutual distance of 100-150 μm.

- Direct electronic image registration by measuring the position of each individual charge pulse with the aid of specially configured read-out electrodes. These electrodes may consist of two orthogonal wire grids where for each grid the individual electrode wires are interconnected by resistors or capacitors. By determining the ratio in signal amplitude at the opposite ends of each wire grid, the position of the center of gravity of the charge cloud can be determined in two dimensions. An alternative technique is the usage of *strip electrodes* of a particular shape, like wedge and strip patterns, that also allows the determination of the center of gravity of a charge cloud by employing a simple position algorithm. It should be noticed that the statistical spread in the determination of these centers of gravity dictates the attainable position resolution. In the case of direct electronic measurement there is practically no loss in image resolution as a consequence of the charge spreading between two plates in the stack.

We shall briefly elaborate here on the four charge centroid read-out techniques that are most commonly used: the crossed grid charge detection (CGCD) technique, the wedge-and-strip technique and the resistive anode technique and the crossed delay line technique.

- The Crossed Grid Charge detection (CGCD) read-out

The crossed grid charge detection system has been applied in X-ray camera's for scientific usage behind grazing incidence optics, notably in space borne instruments on the Einstein, ROSAT and Chandra X-ray observatories. We take these as an example of this read-out technology. The read-out is a crossed grid charge detector (CGCD), which consists of two orthogonal planes of wires electrically separated from each other. The wires are connected to each other by a chain of discrete thin film resistors (R=10 kΩ). The arrangement is depicted at the left side in figure (8.34). Each grid consists of 100 μm diameter gold-plated tungsten alloy wires on 200 μm cen- ters. The two grids are held on a ceramic frame which separates them by 400 μm. At a distance

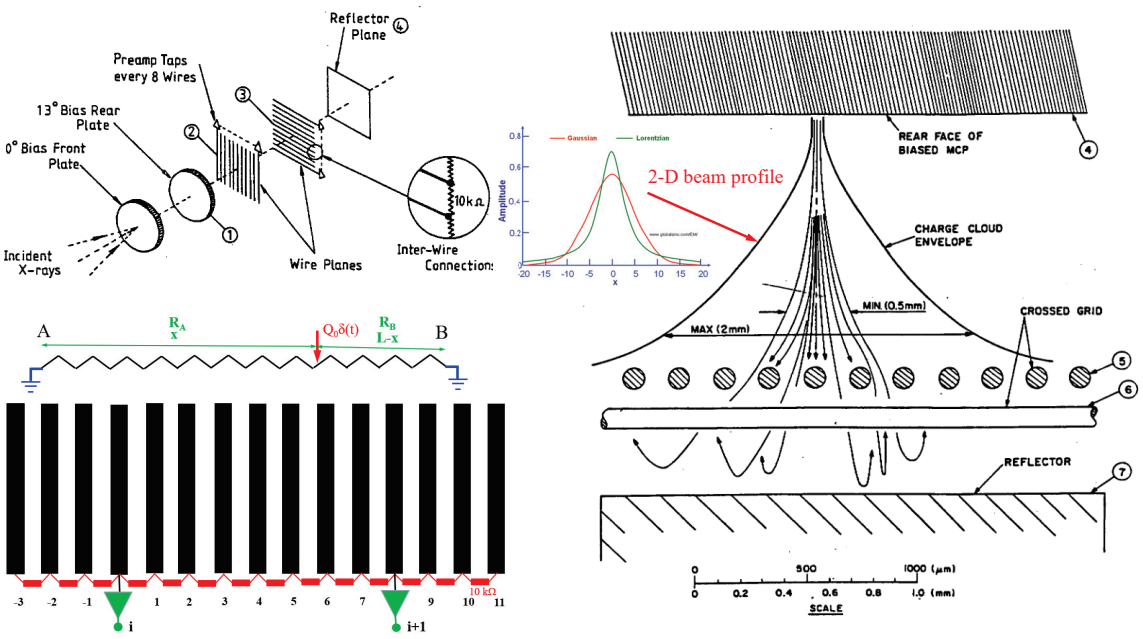


Figure 8.34: *The Crossed Grid Charge Detection read-out system. Upper left: system principal components. Lower left: principle of signal location by resistive charge division, resistively coupled grid wires (10 kΩ) with low-impedance pick-off by charge sensitive (trans-impedance) preamplifiers (electronic taps). Right: spreading of the charge cloud over several wires in the two orthogonal wire planes for charge centroid determination. The wires are evenly spaced with equal open and closed areas. The charge distribution is characterized by denser core with a surrounding halo. Two-dimensionally this is better modeled by a Lorenzian than a Gaussian (see insert).*

of 1000 μm behind the CGCD is a solid reflector plane which is biased about -50V relative to the grids. The grids are biased about +250V relative to the back face of the MCP. Between each wire is a 10 kΩ resistor, every eighth wire is connected to a low input impedance charge sensitive preamplifier shown in figure(8.34) lower left, they act as electronic taps. A 10 x 10 cm read-out would then require approximately 128 preamplifiers for two dimensions.

The position algorithm is based on charge division between adjacent taps when the charge cloud exiting the rear MCP is collected at the wires. As illustrated in figure (8.34) the CGCD is arranged at a distance behind the rear of the MCP stack so that many grid wires do indeed collect charge. If the electron cloud were narrow, the event position would be digitized to the wire spacing. Instead, when the charge cloud spreads over several wires one can exploit the fact that the centroid of the charge cloud can be calculated with a *fine position* algorithm, this centroid can then be determined to a small fraction of a wire spacing. It has been shown from experiment that the electron cloud has a core/halo type of structure, the core has a measured FWHM of 0.65-0.99 mm, or 3 to 5 wire spacings. On the other hand, the halo is large on the scale of the electronic taps. This charge cloud morphology actually implies that a Lorentz distribution (Lorenzian) is better suited than a Gaussian for modeling purposes.

The basic principle of signal location by resistive charge division for a uniformly distributed resistor is depicted at the lower left in figure (8.34). A current impulse $i_0 = \mathbf{Q_0}\delta(t)$ is injected into a uniformly distributed resistor R, at a coordinate x. The currents flowing into the shorted ends A and B are such that:

$$\begin{aligned} i_0 &= i_A + i_B \\ i_A R_A &= i_B R_B \text{ with } R_A, R_B \text{ the resistances between injection and the resistor ends} \\ R_A &= \frac{x}{L} R \\ R_B &= \left(1 - \frac{x}{L}\right) R \quad \text{after some manipulation} \quad \Rightarrow \\ \frac{i_B}{i_0} &= \frac{i_B}{(i_A + i_B)} = \frac{x}{L} \quad \text{or} \quad \frac{\mathbf{Q_B}}{\mathbf{Q_A} + \mathbf{Q_B}} = \frac{x}{L} \end{aligned} \tag{8.36}$$

In case of the wire grids, the situation is slightly different because the resistor is not really continuous but discrete resistances are involved. Figure (8.34) depicts the situation for the wire grids and indicates the positions of two taps, labeled (i) and (i+1). The resistance between two adjacent taps amounts to $8 \,\mathrm{x}\, 10\ \mathrm{k}\Omega$. For this case, a charge deposited on a particular wire between taps (i) and (i+1) will be resistively divided. The fine position x_{fp} (in units of preamplifier tap spacings) is given by:

$$x_{fp} = \frac{\mathbf{Q_{i+1}}}{\mathbf{Q_{i+1}} + \mathbf{Q_i}}, \tag{8.37}$$

past the first tap which collects charge. The total event position is then given by the first tap number (the *coarse* position) plus the the fine position in equation (8.37).
This two-tap algorithm is the simplest form of the coarse/fine position calculation. However for events near a tap when the electron cloud spreads over more than one wire, some of the charge may not be used in the position calculation. This missing charge will have been 'lost' either to tap (i-1) or to tap (i+2). As a consequence this 'charge spill-over' effect will result in distorted positions being calculated. The resulting 'gap' in calculated positions will occur for events at the tap positions. Assessment of the magnitude of these gaps with a simple triangular shaped charge cloud distribution yields a gap size of $\approx 10\%$ of a tap spacing for events located at a tap position. Detailed calculations for charge clouds with different functional forms and characteristic widths show results that are similar to the assessments with the simple triangular case.
One way to avoid the charghe spill-over effect is to make the electron cloud narrow, but as already discussed this leads to a read-out that is digitized at the grid wire spacing. The alternative is to use coarse/fine position algorithms that make use of more than two taps. In the above example of the simple triangular cloud shape, applying a three tap position algorithm:

$$x_{fp} = \frac{\mathbf{Q_{i+1}} - \mathbf{Q_{i-1}}}{\mathbf{Q_{i+1}} + \mathbf{Q_i} + \mathbf{Q_{i-1}}}, \tag{8.38}$$

would completely eliminate the gaps.
This approach can be extended to algorithms that use as many taps as desired in order to insure that there is complete charge collection within the signals used for event position calculation. However one also need to account for the extra uncertainty introduced in

the position calculation due to the electronic noise associated with tha amplifiers and ADC's. In the 'three-tap position algorithm', three ADC's are engaged in detecting an event. The center ADC is that associated with the tap collecting the largest charge (i.e. the coarse position, x_{cp}). The two taps on either side are involved in the determination of the fine position, x_{fp}, that is calculated by using the digitized signals of the amplifiers of three taps:

$$x_{fp} = \frac{\mathbf{S_{i+1}} - \mathbf{S_{i-1}}}{\mathbf{S_{i+1}} + \mathbf{S_i} + \mathbf{S_{i-1}}}, \tag{8.39}$$

where $\mathbf{S_i}$ is the output signal from the ADC of tap (i). It is to be noted that this signal is not exactly commensurate with the charge collected at tap (i), it depends on gain, offset and noise of the associated preamplifier, amplifier and ADC. It can be shown that the noise contributions in the position calculation increase much more rapidly than just the square root of the number of tap signals. This is due to the high weighing of adding additional amplifiers in the position algorithms. Models for the distortions and differential non-linearity that may arise from incomplete charge collection show that these effects can be accurately calibrated *and corrected* if the spill-over is relatively small. The electronic noise considerations associated with the multi-tap fine position algorithms lead to a read-out preference that uses the least number of amplifiers that is still consistent with the charge spill-over constraints. A three tap algorithm with a simple linear correction for charge spill-over distortions is accurate to ≈ 5 μm anywhere between amplifier taps. Position resolutions of ≤ 25 μm over an area of 10 x 10 cm is then readily achievable.

- The wedge-and-strip read-out

A 'blown-up' section of a wedge-and-strip read-out is shown in the left panel of figure (8.35). The anode consists of a quartz plate on which a few μm thick gold layer is evaporated, in which the wedge-and-strip pattern has been photo-etched. The anode surface of the device is divided into three parts representing the wedge, strip and zig-zag electrodes. The triangular wedges increase in width in the vertical direction, whereas the rectangular strips increase in width in the horizontal direction. Hence, position sensing can be accomplished in the vertical and horizontal direction by the wedges and the strips respectively. In this particular example, the minimum and maximum wedge or strip width range from approximately 50 μm to 350 μm, the linear dimension of both wedges and strips and of the anode plate can amount to several centimeters. The zig-zag electrode occupies the remaining surface area. The insulating gaps between the electrodes, also indicated in figure (8.35) have a width of typically 50 μm. The pitch of the wedges and strips on the anode, i.e. the fixed distance between the centers of two adjacent wedges (or strips), has to be selected in relation to the size of the charge cloud that arrives at the anode. To prevent non-linearities in the position determination due to undersampling of the size of the charge cloud, this pitch should be at least three times smaller than the charge spot on the anode (Nyquist's sampling criterion). The **x** and **y** positions of the centroid of the electron cloud emanating from the MCP are then proportional the the fractions f_s and f_w of the total charge deposit on the anode:

$$f_s = \frac{\mathbf{Q_s}}{\mathbf{Q_s} + \mathbf{Q_w} + \mathbf{Q_z}} \quad \text{and} \quad f_w = \frac{\mathbf{Q_w}}{\mathbf{Q_s} + \mathbf{Q_w} + \mathbf{Q_z}}, \tag{8.40}$$

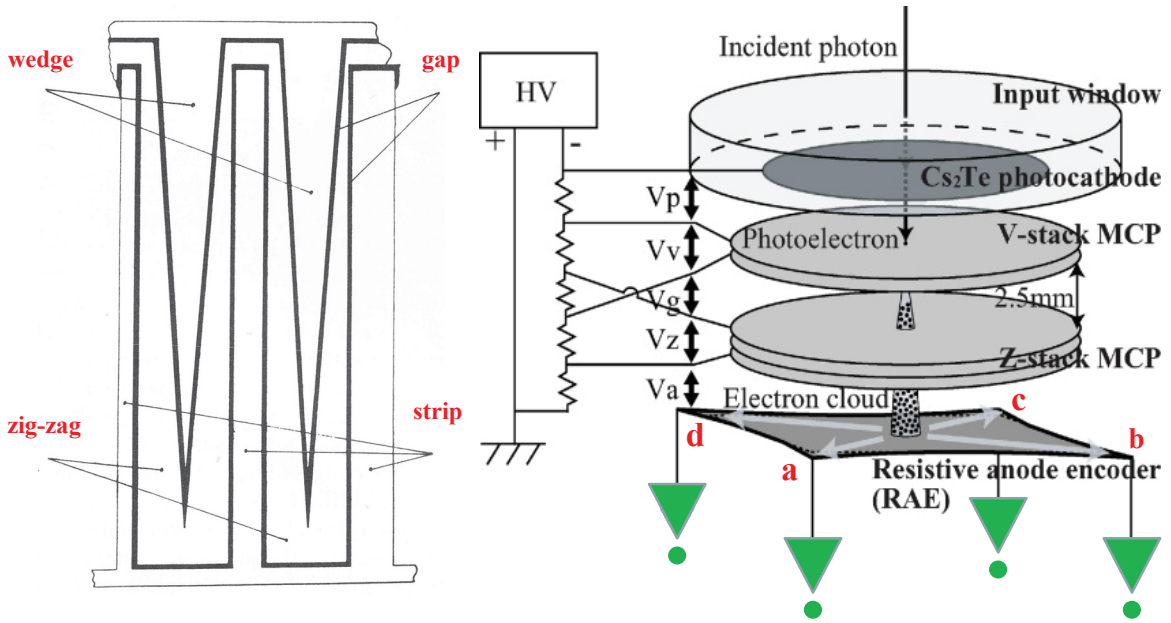


Figure 8.35: *Charge cloud read-out techniques: (left) wedge and strip read-out. (right) delay line read-out.*

where $\mathbf{Q_s}$, $\mathbf{Q_w}$, and $\mathbf{Q_z}$ are the charge signals on the s-, w-, and z-electrodes respectively.

- The resistive anode read-out

One of the most extensively used photon counting read-out schemes is the resistive anode. The anode consists of a uniform resistivity coating (several 100 kΩ), applied to an insulating substrate. A simple two dimensional read-out can be a square sheet with contacts at the corners, mounted close (< 1 mm) behind the MCP output face, see right panel in figure (8.35). The diffusion of charge from the event location gives pulse amplitudes and pulse rise times proportional to the distance from the contact. The spreading of charge can be described by the transport along a two-dimensional transmission line (i.e. a RC-line) and is governed by the diffusion equation according to:

$$\frac{\partial V}{\partial t} = \frac{1}{RC}\left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2}\right) \quad \text{with } R \text{ and } C \text{ normalized per square,} \tag{8.41}$$

Event positions may now be determined from charge division between opposing contacts. In each case only four amplifiers are required for full two-dimensional imaging. In the charge division mode the centroid location of the photon impact point P_x along the x-axis and P_y along the y-axis is given by:

$$P_x = \frac{\mathbf{Q_a}}{\mathbf{Q_a} + \mathbf{Q_c}} \quad \text{and} \quad P_y = \frac{\mathbf{Q_b}}{\mathbf{Q_b} + \mathbf{Q_d}}, \tag{8.42}$$

where $\mathbf{Q_a}(\mathbf{Q_b})$ and $\mathbf{Q_c}(\mathbf{Q_d})$ are the charge signals on two opposing anode corners. Alternatively, analysis of the pulse profile also provides the position information. Digitization

of the contact signals with a resolution that is at least twice the resolution of the anode provides the necessary oversampling (Nyquist criterion) to retain the anode resolution. simple two-dimensional resistive anodes display considerable image distortion due to charge reflection effects since perfect termination with the characteristic impedance of a two-dimensional transmission line is not feasible.This problem can be alleviated by using a resistive anode with low resistivity borders.The position resolution of resistive anode sensors is determined by the the thermal (Johnson) noise of the resistance layer. High-gain stacked MCP configurations are commonly used in combination with resistive anodes to produce a high signal to noise ratio . Resolutions of 50 μm FWHM over a few centimeters have been achieved in this way.

- The delay line read-out

When a charge pulse leaving the MCP back face impinges at a specific position on a

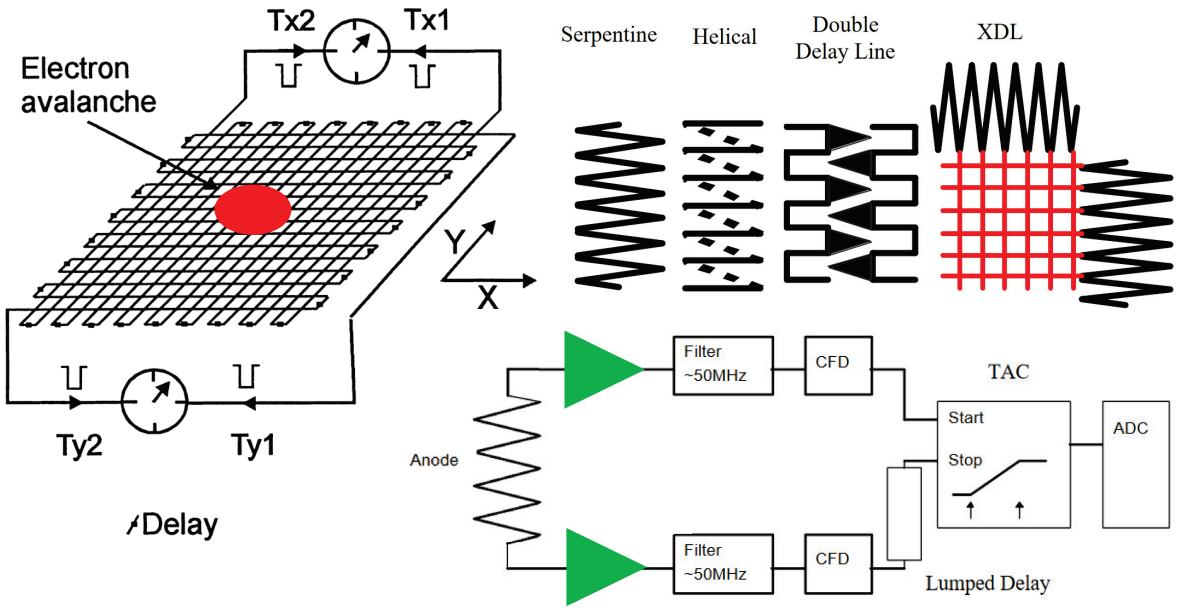


Figure 8.36: *Charge cloud read-out techniques: the delay line read-out.*

delay line anode, the pulse splits in two parts and travels in both directions along the delay line. The time difference between the time of arrival of each pulse at both ends of the anode is linearly proportional to the original position on the anode and inversely proportional to the velocity of propagation of the pulses in that coordinate. This is schematically shown in figure (8.36).

To minimize the error in the determination of the position (i.e. optimize spatial resolution), one must decrease the velocity of propagation as much as possible and minimize the timing error of both pulses. The geometry, layout and materials of the delay line determine the velocity of the pulses while the timing error is usually determined by the pulse sensing electronics. The two are not completely independent, however, as certain choices of very slow delay line patterns can adversely affect the pulse shape which in turn increase the timing error. Typical delay line velocities are of order 1 mm per

nanosecond, while output pulses are about 5 nanoseconds wide. Therefore to achieve a spatial accuracy of 25 μm requires that the measurement of the arrival time of this 5 nanosecond pulse be accurate to 25 picoseconds. An assortment of delay line patterns are available, we mention here specifically here the *serpentine* and *helical* delay lines, they refer to single or multi-planar delay line patterns. When used as actual anodes, they are usually connected to conducting 'strips' so that the delay line is separated from the active charge collection region, see figure (8.36). Examples of this are: the 'Double Delay Line' or DDL where the strips are wedges and the orthogonal coordinate is determined by a charge division scheme between the wedges; and the Cross Delay Line or XDL, where two orthogonal sets of fingers cover the active charge collection region and each set is connected to a delay line (of either type). The choice of anode pattern is usually based on the size of the detector, the resolution requirement in each dimension, and the cost and reliability of the manufacturing techniques. The impact of delay line choice on the timing electronics is the effect that it has on the pulse shape. Both helical and serpentine delay lines are not perfect transmission lines and do not have a linear phase characteristic essential to avoid pulse dispersion, overshoot and ringing. Pulses widen as they traverse the delay line and lose amplitude. Most of this is due to capacitive and inductive coupling along the delay line. When unterminated strips are attached at every bend in the delay line, as in the DDL or XDL anodes, a fraction of each pulse must travel down these strips to be reflected back into the pulse but at a later time dependent on strip length. These strips act as resonance filters and tend to slow the velocity and widen the pulse.
A simple schematic of the one dimensional delay line readout electronics is also shown in figure (8.36). Pulses from each side of the anode (*approx* 10 mV, 5 nanoseconds wide) are amplified, high and low pass filtered and then sent to a Constant Fraction Discriminator (CFD). The CFDs produce the ECL logic pulses to start and stop a Time to Amplitude Converter (TAC) which outputs an analog voltage pulse whose amplitude is proportional to the time difference of the start and stop signals. This pulse is then digitized and represents the pixel location of the input pulse. Since the delay lines are usually symmetric, the start and stop designations are arbitrary until an extra delay is added on the stop side to ensure that start always comes before stop for a given pulse. The added delay on the stop side can either be before or after the CFD. Usually the delay is added after the CFD with a compact lumped-delay circuit. However, the lumped delay can degrade the pulse shape, however this is not a real concern for large ECL pulses. The alternative is to add an extra length of low dispersion cable to the input to the stop CFD, but this is a very bulky solution when the delays of the anode are of order a hundred nanoseconds.

8.3.6 Temporal response, recovery time

The transit time of electrons in a MCP depends linearly on the channel length and the transit time jitter is proportional to the transit time for a given plate voltage and the length-to-diameter ratio α. MCP's with channel diameters less than 12 μm produce charge pulses having a FWHM of less than 1 nanosecond, e.g. a chevron MCP with a channel of 5 μm produces a charge pulse width of $\approx$ 500 picoseconds with a time spread

of only 60 picoseconds. In reality this time spread may actually become dominated by noise in the read-out electronics with an increase to ≈ 150 picoseconds FWHM. This gives the MCP unique capabilities for fast timing applications, including time of flight measurements in particle beams and in mass spectrometry. They are at last one order of magnitude faster than photomultiplier tubes that produce pulses having a FWHM in the range from 5-30 nanoseconds.
A potential disadvantage of MCP's is their relatively slow recovery time. When a charge pulse has been produced, a positive charge equal to the delivered electron charge remains near the exit of the MCP. This positive charge generates an additional electric field that suppresses subsequent electron amplification. This additional field needs to be neutralized by the plate supply current, however this takes time due to the high plate resistance. The *dead time* per MCP-channel can be approximated by:

$$T_d = \frac{\mathbf{Q}_{out}}{i_{ch}} = \frac{qGNR}{nV}, \tag{8.43}$$

with G the MCP gain, N the total number of MCP channels, R the plate resistance, V the plate high voltage and n the number of channels in the rear plate that are triggered by a 'single electron event' at the entrance of the front plate. Substituting nominal operational values like G=10^6, V= 1 kilo-Volt and R= $10^9\,\Omega$ we get $T_d \approx 20$ msec, implying a gain drop of a factor 2 at 50 Hz.

8.3.7 Photocathodes and Image intensifiers

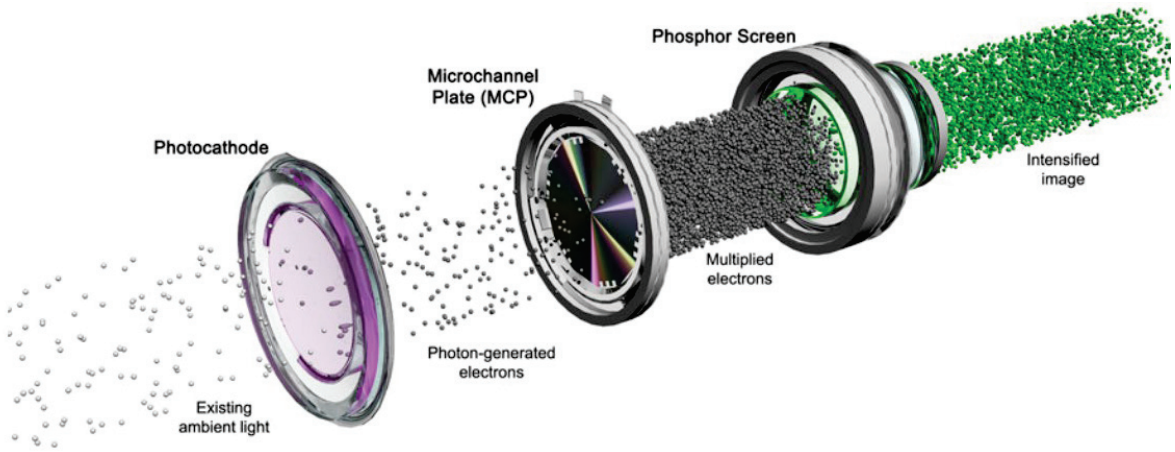


Figure 8.37: *Principle of an image intensifier tube employing an MCP as amplifying element.*

Today the major emphasis in terms of imaging devices is on CCD's and CMOS arrays. Yet the unique capabilities of MCP's and MCP stacks, in particular their vast use in night viewing applications, image intensifiers and particle detection suggest that high quality lead glass MCPs will still be around for a long time to come.

For radiation detection the MCP needs to be equipped with a photocathode that provides the maximum efficiency in the wavelength range under consideration, these

photocathodes are coated onto the front face of the MCP(-stack). In terms of quantum efficiency the same holds as for photomultiplier tubes, i.e the desirability of negative electron affinity cathodes whereby in the case of the MCP, the angle of incidence of the incoming photon is also of major importance for the achievable quantum efficiency. The most important photocathode materials comprise Lithium-Fluoride (LiF), Magnesium-Fluoride (MgF_2) and Cesium-Iodide (CsI).
In the case of MCP-amplified image intensifiers the photocathode is physically separated from the MCP, it is free standing in the sense that the photocatode material is coated onto the entrance window, i.e. the objective lens of the intensifier. An image intensifier tube is a device that intensifies (or amplifies) low light level images to levels that can be seen with the human eye or detected by digital image sensors. All modern tubes consist of three main components, a photocathode, a MCP and a phosphor screen. The principle is shown in figure (8.37). Image intensifier tubes collect the existing ambient light through the objective lens of a night vision device. The light may originate from natural sources, such as starlight or moonlight, or from artificial sources

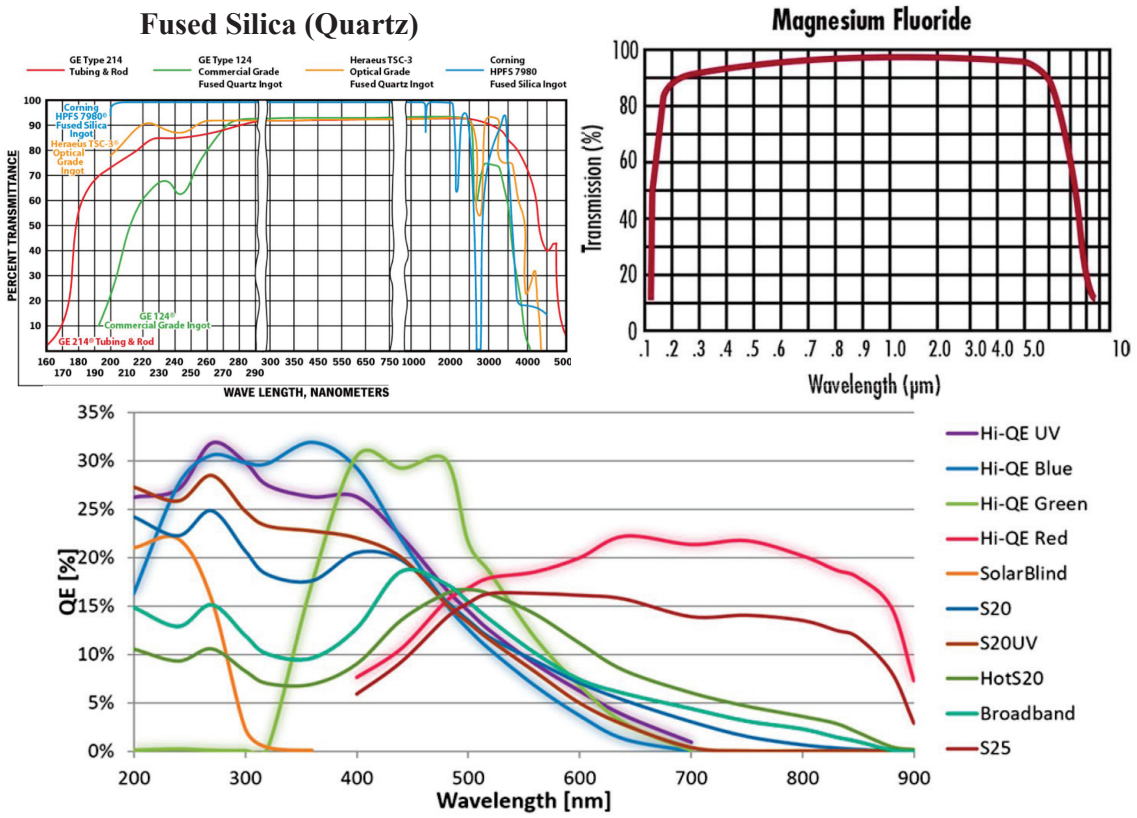


Figure 8.38: *The broad band transmission characteristics of fused silica (quartz) as a function of wavelength, extending from 180-4500 nanometers, i.e. from the vacuum UV into the near infrared. If needed an even larger coverage, to both shorter UV (including the Hydrogen-Lyman-α line at 121 nm) and to longer waves in the infrared, can be achieved with MgF_2, see the upper right panel. The lower panel shows quantum efficiencies for a variety of photocathode materials that are commercially incorporated in image intensifier tubes.*

such as streetlights or infrared radiation sources. The low level of incoming light enters the tube through its input window and strikes the photocathode.
The photocathode is a very thin light sensitive layer deposited on the inside of the input window that converts the photons into electrons and releases them into the vacuum of the tube. All modern intensifier tubes operate under a vacuum of about 10^{-9} to 10^{-10} Torr, which is essential to protect the photocathode from oxidation and rapid destruction. Once released by the photocathode, the photo-electrons are accelerated and focused by a high electrical field towards the MCP. When applying a single MCP, for each electron that enters the MCP, approximately 10^3 to 10^4 electrons are generated and subsequently accelerated from the output of the MCP by a third electrical field towards the phosphor screen. The phosphor screen is a thin phosphorous light emitting layer deposited on the inside of the output window of the intensifier tube, that usually comprises a fiber optics assembly, that converts the electrons back into photons. When the multiplied flow of electrons out of the MCP strike that layer, tens of thousands of photons will be generated for every single photon that was initially converted by the photocathode. This entire multistage process creates an "intensified" image, much brighter than the original image, which can subsequently be seen by the human eye or a digital image sensor like a CCD or CMOS camera.
It is clear from the photocathode quantum efficiencies displayed in figure (8.38) that the sensitivity for EM-radiation reaches into the short wavelength infrared (750-950 nanometer). Without an additional light source, intensified images can only be produced as a result of the presence of some residual visible light. Presently, second and third generation intensifiers (I2 and I3) can be equipped with so-called *infrared illuminators* that emit short wave infrared light at wavelengths between 850 and 950 nanometer, invisible for the human eye, that reflects of the local scenery and gets subsequently amplified by the intensifier tube. In summary, image intensifiers primarily operate with reflected light. Alternatively, night vision can be prominently accomplished with the aid of *contrast in the infrared emissivity* of the ambient due to thermal gradients introduced by the presence of a large variety of *heat sources*. This has now evolved in the powerful technique of thermal imaging that detects and maps the distribution of heat radiation by scans in the infrared.

8.4 Thermal imagers

8.4.1 Heat sensing, wavelength domain

The thermal imager system's function is to produce a picture that is a map of temperature differences related to spatial flux and emissivity differences across an extended target, i.e. the thermal image is a visual capture of heat. The more heat an object emits, the brighter it appears in a thermal image. Thermal images give the most information when there are significant temperature differences in a scene.
The most common presentation of thermal images is in black, white, and gray. The different temperature values are then translated into 256 grayscale values. The most common presentation (or palette) is white-hot, in which heat sources appear white against lower-temperature gray and black backgrounds. In some cases, black-hot, where sources

of heat appear black, may be easier to use. Most cameras can switch between palettes. Thermal images are sometimes associated with bright, intense colors, which may seem a little odd considering that the camera works outside the spectrum of visible light. Since the human eye is better at distinguishing different shades of color than different shades of gray, adding color sometimes makes it easier to see differences in thermal images. These so-called pseudocolors are created digitally each color or nuance represents a different temperature, usually ranging from white and red for higher temperatures to green, blue, and violet for colder ones. Figure (8.39) shows an example: the thermal image in pseudo-color of a hot air balloon

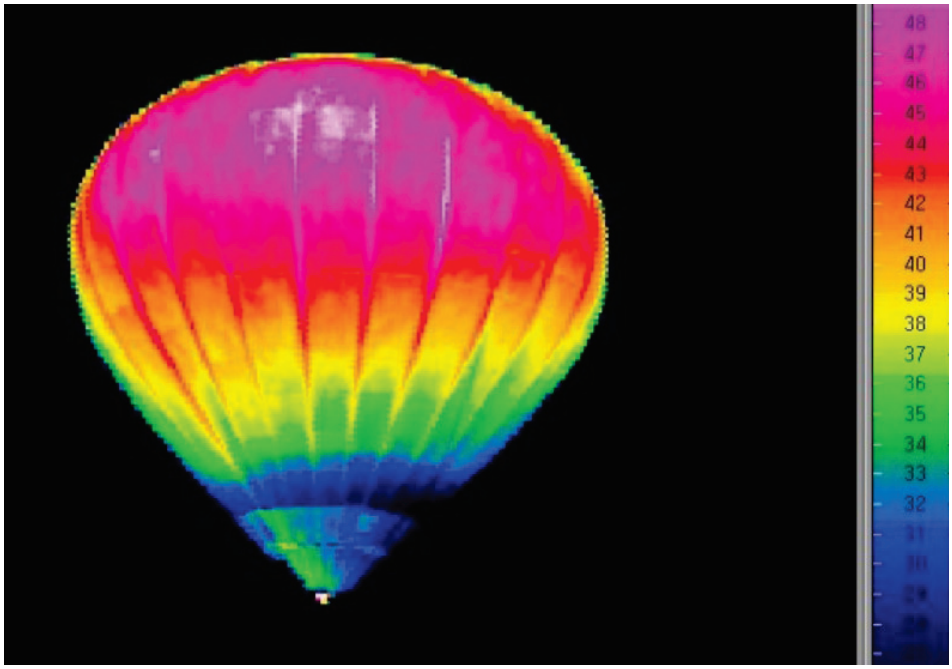


Figure 8.39: *Pseudo color thermal image of a hot air balloon.*

The capture of heat occurs in the infrared wavelength domain. This infrared waveband is subdivided in a number of windows with increasing wavelength, i.e. short-wave infrared (SWIR, 1.4−3 μm), mid-wave infrared (MWIR, 3−8 μm), long-wave infrared (LWIR, 8−15 μm), and far-wave infrared (FWIR, 15−1000 μm). Earth scientists define the LWIR wave band as thermal IR, but in the industry procuring thermal imagers the MWIR wave band is also commonly referred to as thermal. However, the 5−8 μm part of the MWIR wave band is virtually unusable for thermal imaging purposes because of the high spectral absorption of the atmosphere in this range.

The thermal imaging industry often divides the electromagnetic spectrum based on the response of various IR detectors. Very long-wave infrared (VLWIR) is then inserted between LWIR and FWIR, and also the boundaries between the other ranges are sometimes defined slightly different. Figure (8.40) shows the positions of these IR sub-bands as part of the full electromagnetic spectrum and, in addition, the transmission of the atmosphere at ground level From the latter it is clear that for thermal imaging the MWIR band effectively runs from 3−5 μm. The main thermal window is in the LWIR

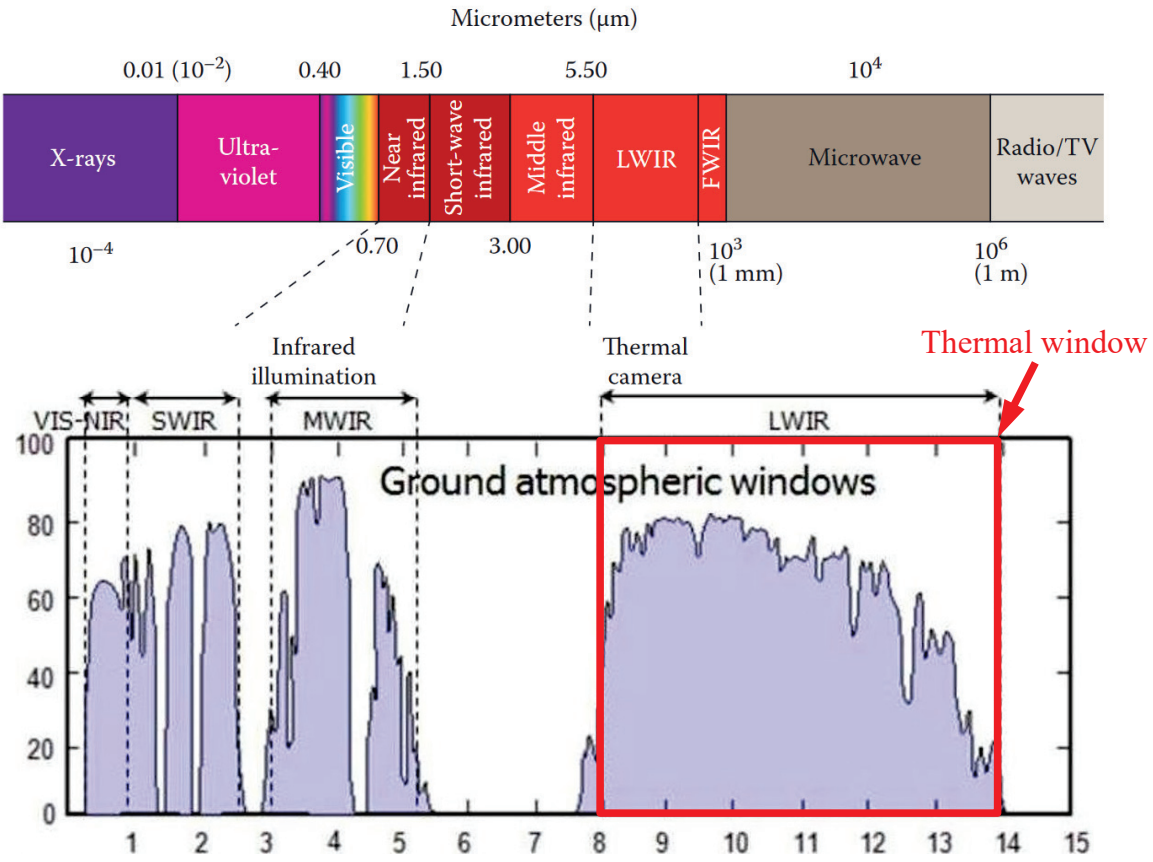


Figure 8.40: *Nomenclature for waveband identification in the electromagnetic spectrum from radio waves to X-ray photons. The atmospheric window in the thermal infrared lies in the LWIR window and ranges from approximately 8 to 14 μm.*

band effectively ranging from 8−14 μm indicated by the the red-lined box.
The camera's sensitivity can be defined as its capability to distinguish between temperature differentials. The greater the temperature difference in a scene, the clearer the thermal images will be. Of course the contrasts in a thermal image also depend of on the emissivity of its objects.

8.4.2 Sensors in a thermal camera

The sensor in a thermal camera is a 2D-array of thousands of small sensor elements (pixels) that are sensitive to thermal IR radiation. Sensors used for thermal imaging can be broadly divided into two types: cooled and uncooled IR-sensors. For the thermal window considered here, which contains the peak black body emissivities between 250 and 450 K, uncooled IR-sensors are the imagers of choice and hence we shall limit the discussion here to these devices. An uncooled IR sensor is stabilized at or close to the ambient temperature, using relatively simple temperature control elements or no temperature control at all. Sensors of this kind operate in the LWIR band.
A common design is based on micro-bolometer technology, a schematic overview of

which is shown in figure (8.41). The bolometer comprises an array of thin plates that 'float' above the surface of a Read Out Integrated Circuit (ROIC), e.g. a CMOS imager. The tiny plates constitute (multilayer)Vanadium Oxide (V_xO_y) or amorphous silicon (α-Si) thin films with a large resistance temperature coefficient on a silicon element with large surface area, low heat capacity and good thermal isolation. When thermal IR-radiation from a specific range of wavelengths hits this material, the electrical resistance changes. The thin plates can be regarded as micro-bridges that are connected via electrodes to the input cells of the read-out integrated circuit (ROIC), e.g. a CMOS array, see the schematic pictures in figure (8.41). Originally the advantage of α-Si was that those uncooled sensors could be fabricated in a silicon foundry. Presently Vanadium Oxide coated sensing elements can also be manufactured in a silicon foundry, which has led to a dominance of this material in present day uncooled IR-sensor technology. The individual elements in an uncooled sensor array can respond in slightly different ways to the incoming IR-radiation, which would result in 'drift' in individual pixel values. To remedy this, the sensor performs a so-called non-uniformity correction. A mechanical shutter blocks the sensor and provides it with a standard temperature target, against which every pixel is corrected. This process occurs at regular intervals or when a specific temperature change takes place.
Another kind of micro-bolometer is based on ferro-electric (i.e. pyro-electric) technology. Here, small changes in the temperature of the sensing material create large changes in electrical polarization, the physics of which we have already covered in the section on pyroelectric sensors. Ferro-electric material for microbolometers involves mostly Barium Strontium Titanate (BST).
The diagram at the lower left in figure (8.41) shows the fractional percentage of these three technologies in the current thermal camera commercial market, showing the dominance of the V_xO_y thin film application.

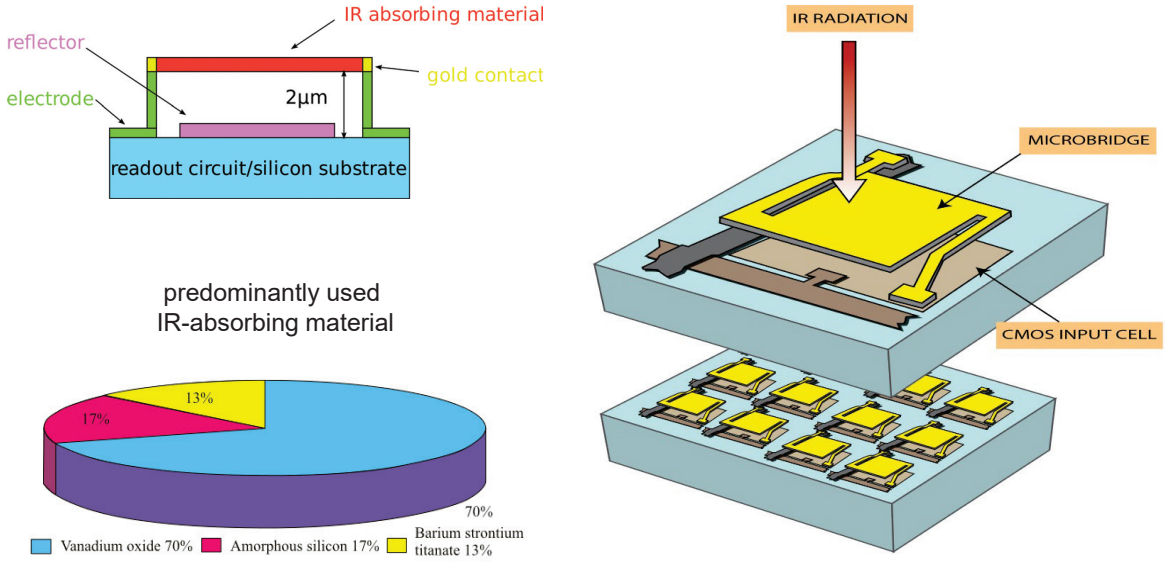


Figure 8.41: *Elements of a thermal imager.*

8.4.3 The Noise Equivalent Temperature Difference (NETD)

The camera's sensitivity to thermal radiation, which determines its ability to distinguish different temperature differences in a scene, can be expressed as its noise equivalent temperature difference (NETD) value. We shall now elaborate on this important figure of merit for thermal camera's by deriving a quantitative expression for the NETD.
Figure (8.42), left panel, shows the basic configuration of a thermal imaging system. The radiance L at the scene location filling the instantaneous field of view (Ω_{pix}), produces an energy flux ϕ_s on the sensor element:

$$\phi_s = L \cdot A_{ap} \cdot \Omega_{pix} \Rightarrow \Omega_{pix} = \frac{A_s}{f^2} = \frac{A_{footprint}}{R^2} \Rightarrow \phi_s = L \cdot \frac{\pi}{4} D_{ap}^2 \cdot \frac{A_s}{f^2} \tag{8.44}$$

where A_{ap}, D_{ap} are the area/diameter of the camera aperture, A_s, Ω_{pix} area of the sensor element/solid angle of the scene footprint $A_{footprint}$ and R, f the camera-object distance (range)/camera focal length respectively. From this we get:

$$\phi_s = L \frac{\pi}{4} \frac{A_s}{(f/\#)^2}, \quad \text{with } f/\# \text{ the focal number of the camera} \tag{8.45}$$

From this equation we see that the irradiance ϕ/A_s on the sensor is independent of

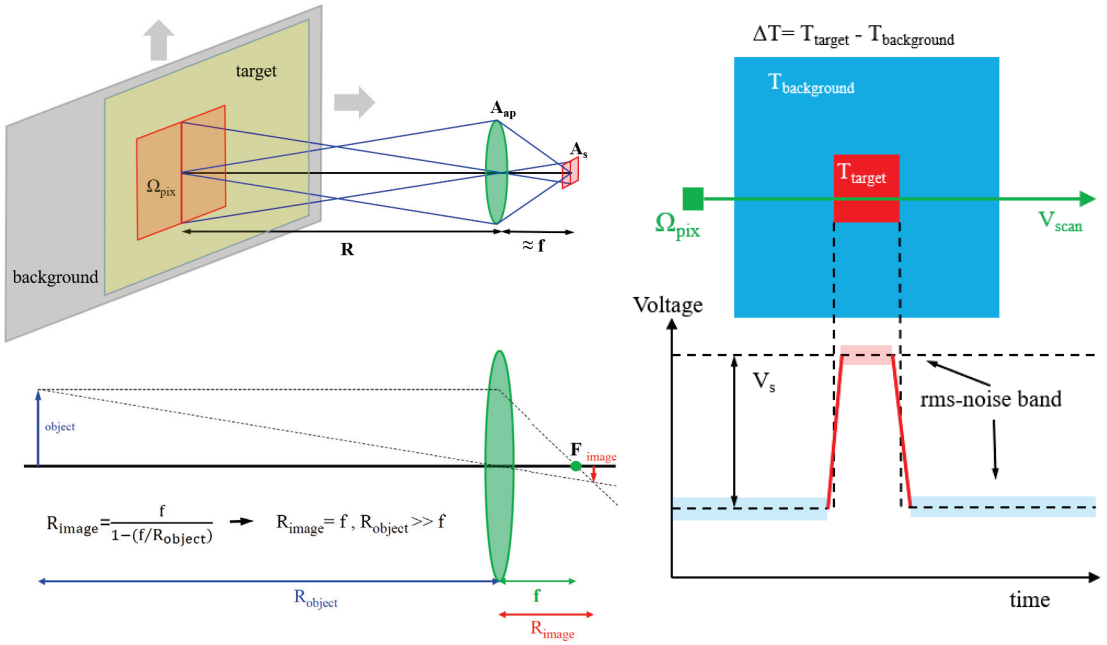


Figure 8.42: *(Upper left panel) Thermal Imager system configuration. (Lower left panel) Relation between object distance (R_{object}), image distance (R_{image}) and camera focal length (f). (Right panel) NETD calibration set-up and the resulting output signal profile.*

the range R and only depends on the source radiance and the focal ratio of the camera. Note that we have implicitly assumed that the extended heat source completely fills

(or overfills) the pixel solid angle Ω_{pix}. We can now obtain an expression for the signal voltage V_s produced by the sensor in response to the heat signal by multiplying equation (8.45) by the voltage responsivity R_V of the sensor:

$$V_s = R_V \cdot L \frac{\pi}{4} \frac{A_s}{(f/\#)^2} \tag{8.46}$$

We should of course note that the sensor responsivity and the source radiance are functions of wavelength and that their product in equation (8.46) denotes an integral over the wavelength passband of the camera system:

$$R_V \cdot L = \int_{\lambda_1}^{\lambda_2} R_V(\lambda) L(\lambda) d\lambda \tag{8.47}$$

Since we are interested in the thermal *mapping* of the heat source, we have to write an expression for the *change of the signal voltage* for a *change in temperature* by taking the partial derivative $\partial/\partial T$ of each side of equation (8.46):

$$\frac{\partial V_s}{\partial T} = R_V \cdot \frac{\partial L}{\partial T} \frac{\pi}{4} \frac{A_s}{(f/\#)^2} \tag{8.48}$$

Substituting for $R_V = (D^* \sqrt{\overline{V_n^2}})/(\sqrt{A_s \Delta\nu})$ and after some rearrangement we can write for the signal to noise ratio $SNR = \Delta V_s / \sqrt{\overline{V_n^2}}$:

$$SNR = \Delta T \cdot \frac{D^*}{\sqrt{\Delta\nu}} \cdot \frac{\partial L}{\partial T} \cdot \frac{\pi}{4} \frac{\sqrt{A_s}}{(f/\#)^2} \tag{8.49}$$

Equation (8.49) can now be written in terms of the $NETD$, which is the value of the temperature difference that produces unity SNR. Setting $SNR = 1$ in (8.49) and solving for ΔT we find for the $NETD$:

$$NETD = \frac{4}{\pi} \left[\frac{(f/\#)^2 \sqrt{\Delta\nu}}{D^* (\partial L/\partial T) \sqrt{A_s}} \right] \tag{8.50}$$

The $NETD$ characterizes the thermal sensitivity of an infrared sensor system. A smaller $NETD$ indicates a better thermal sensitivity.
For the best sensitivety, i.e. the lowest $NETD$, we should maximize the spectral integral of the product of D^* and the radiance contrast $\partial L/\partial T$:

$$\int_{\lambda_1}^{\lambda_2} \frac{\partial L}{\partial T}(\lambda, T) \cdot D^*(\lambda) d\lambda \quad \Rightarrow \quad \text{maximized} \tag{8.51}$$

This will be approximately the case when the peak of the spectral responsivity ($\propto D^*$) and the peak of the spectral radiance contrast coincide.

Note: The wavelength of maximum radiance is not the same as the wavelength of maximum radiance contrast, since we have:

$$\frac{\partial L(\lambda,T)}{\partial \lambda} = 0 \Rightarrow \lambda_{max} = \frac{2898\,\mu m K}{T} \Rightarrow = 9.7 \mu m \text{ at } T = 300\,K \quad (8.52)$$

$$\frac{\partial}{\partial \lambda}\left[\frac{\partial L(\lambda,T)}{\partial T}\right] = 0 \Rightarrow \lambda_{max} = \frac{2410\,\mu m K}{T} \Rightarrow \approx 8\,\mu m \text{ at } T = 300\,K \quad (8.53)$$

Note: The numerical values quoted are derived for a black body spectral radiance distribution.
For maximum thermal sensitivity, the passband of a thermal imager system should include the peak of the radiance contrast. However it may not always be possible to satisfy these conditions because of other constraints, such as atmospheric obscurance or available sensor technologies.
Equation (8.50) also shows that the $NETD$ is proportional to the square root of the frequency bandwidth, this is intuitive since the rms-noise is proportional to the square root of the frequency bandwidth. The equation also implies that the $NETD$ becomes lower with a lower focal ratio. This immediately follows from equation (8.45): a lower focal ratioa results in more energy flux captured by the sensor element. A larger detector area also lowers the $NETD$ value, i.e $\propto 1/\sqrt{A_s}$. This is a critical parameter since a system level descriptor dictates that while the thermal sensitivity of the imager improves with larger A_s, the spatial resolution becomes poorer for larger sensor pixels. Hence the $NETD$ is, when considered in isolation, not a suitable design parameter. In fact, the NETD is a purely radiometric parameter which ignores the presence of diffraction effects. It implicitly assumes that the Optical Transfer Function (OTF) of the camera, which magnitude is represented by the diffraction Modulation Transfer Function (MTF) of the lens system, yields an impulse angular response that is by definition smaller than the solid angle subtended by the sensor element. This then allows a geometrical-optics calculation of Ω_{pix} which may violate the cameras OTF. Another parameter, the Minimum Resolvable Temperature Difference (MRTD), takes both thermal sensitivity and spatial resolution into account and is therefore more suitable for design purposes. We shall return to the signature of this parameter shortly.
In practise, actual measurement of the $NETD$ calibrates the thermal sensitivity of the imager system. The calibration set-up is also shown in figure (8.42) and features a region of higher temperature superimposed on a background of similar but lower temperature. The high temperature field should overfill the instantaneous field of view Ω_{pix} of the sensor element and the temperature difference between the high temperature field and the background should be several times the expected $NETD$ to ensure that that the sensor response is easily visible above the system noise The spectral distributions of the target and the background regions normally approximate sufficiently well blackbody distributions. From the sensor scan over the temperature distribution, displayed in figure (8.42), the $NETD$ can be calculated as the temperature difference that would produce a SNR of 1:

$$NETD = \frac{\Delta T}{(V_s/\sqrt{\overline{V_n^2}})} \quad (8.54)$$

If for example a SNR=50 is obtained for a temperature difference $\Delta T = 2$K, the NEDT equals 40 mK. Most thermal network cameras have an $NETD$ value of 50-100 mK, though there are newer generations of bolometers that have an $NETD$ as low as 20 mK.

8.4.4 Resolving an image: the Minimum Resolvable Temperature Difference

Because regular glass blocks thermal radiation, regular glass–based optics and lenses cannot be used in thermal cameras. Currently, *Germanium* is the most commonly used material for thermal camera optics. This very expensive metalloid, which is chemically similar to tin and silicon, blocks visible light while letting through the IR light. Not all lenses are pure Germanium. For example, some are made of a Germanium-based material called *chalcogenide glass*, which allows a wider spectrum of IR-light to pass through. Like with most materials, there are benefits and disadvantages. Chalcogenide glass contains cheaper materials and is moldable. However, the master mold requires a significant initial investment that can only be justified when producing at large quantities.
Resolutions are generally much lower for thermal cameras than for conventional network cameras. This is mostly due to the more expensive sensor technology involved in thermal imaging. The pixels are larger, which affects the sensor size and the cost of materials and production. Currently, typical resolutions for thermal cameras range from 160 x 128 to high resolutions of 640 x 480 (VGA), though even higher resolutions are available.
In thermal imager systems, a combination of both spatial resolution and thermal sensitivity determines the final performance of the system. In observing objects with low spatial frequencies, thermal sensitivity is most important, whereas for targets with high spatial frequencies, spatial resolution is the dominant aspect. The Minimum Resolvable Temperature Difference (MRTD) *figure of merit criterion* combines thermal sensitivity and spatial resolution. The MRTD is an overall measure of performance, but also constitutes a *design criterion*. It basically answers the question: what temperature difference is required for various-sized 'bar' targets (featuring bars(stops) and open spaces) providing a range of fundamental spatial frequencies ξ_b to be visible on a display? Smaller is better for MRTD, just as in the case for the NETD. The spatial resolution requirement for a thermal imager can be assessed by consideration of the so-called *Johnson criterion* that provides a way to describe real targets in terms of simple visual patterns, i.e. square wave patterns with a selection of bar widths (equivalent of black/white line pairs with a selection of pitches). This was all pioneered by the military who analyzed the performance of an average observer in executing a certain decision task in terms of the number of just resolved bar cycles that would fit across the minimum dimension of the target. This approach for assessing the required target resolution is illustrated in figure (8.43). More complex decisions require a higher level of detail and thus better image resolution. *Detection*, i.e. presence discerned, was found to require *one cycle* per minimum dimension, *Recognition*, i.e. tractor versus truck versus tank, required *4 cycles* per minimum target dimension, while *Identification*, i.e. a particular specimen,

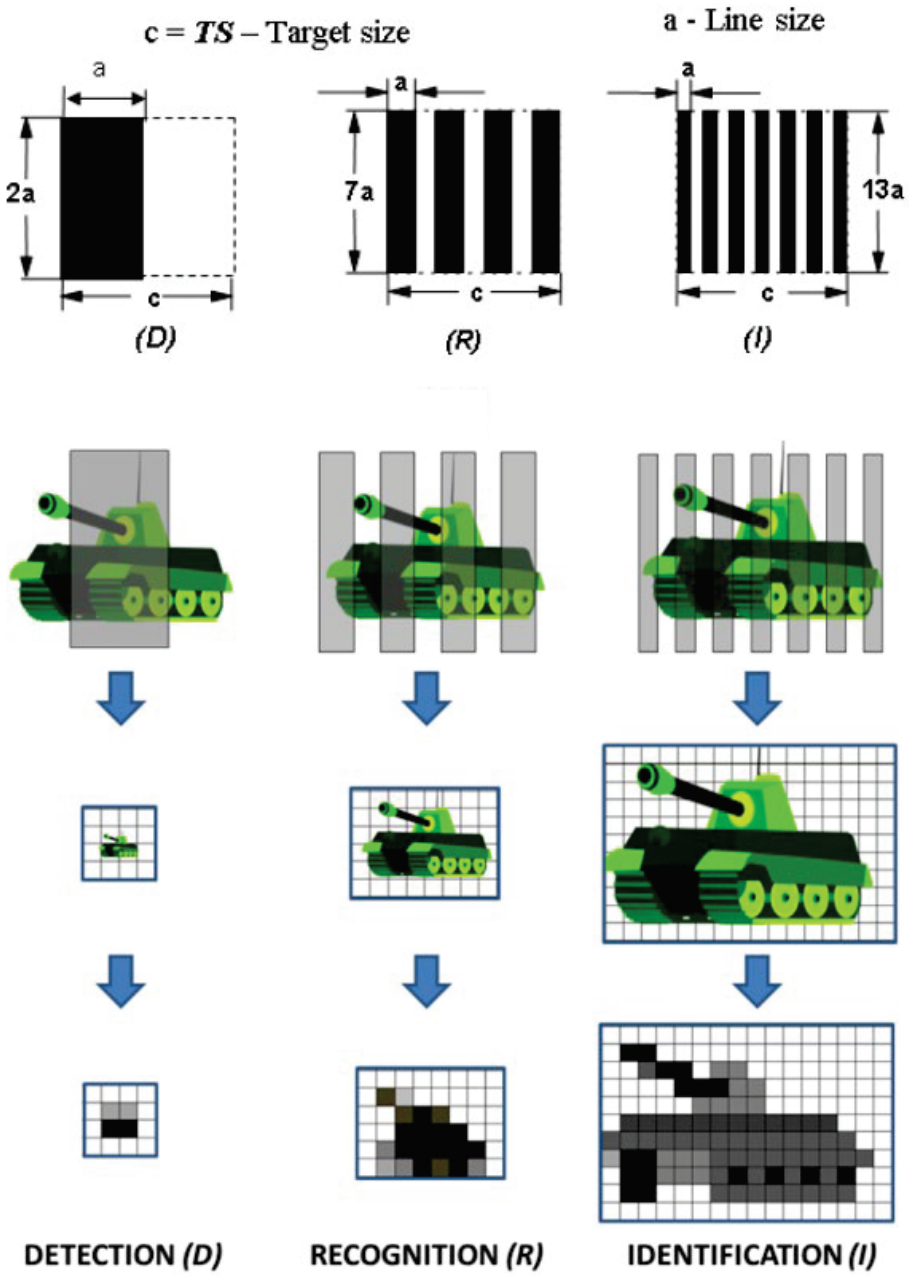


Figure 8.43: *Assessing the required resolving power of a thermal imager with the number of bar cycles across the minimum dimension of a target necessary for detection, recognition and identification respectively.*

required *6.4 cycles* per minimum target dimension. These resolutions allow a 50% probability of correct detection, recognition and identification respectively. A higher level of correctness would require finer resolutions, i.e. a higher number of cycles, then those just quoted. So, to realize the desired decision task, it should fit with the relationship between target size, target distance, camera focal length and imaging sensor dimension. If the required number of cycles per minimum target dimension is determined from the Johnson criterion (n_{cycle}), the required *angular spatial frequency* ξ in *cycles/radian* is

given by:

$$\xi = \frac{n_{cycle}}{(w/R)}, \quad w = \text{minimum target dimension, } R = \text{distance to the target} \tag{8.55}$$

The denominator in the above equation is the angular subtense of the target as seen from the thermal camera. The Nyquist sample criterion dictates *two pixels per cycle* for the highest angular spatial frequency that is commensurate with the decision task. This determines the linear angular pixel size (α_{pix}) of the instantaneous field of view (Ω_{pix}), $\sqrt{A_s}$ represents the linear pixel pitch:

$$\xi = \frac{n_{cycle}}{(w/R)} = \frac{1}{2\alpha_{pix}} = \frac{1}{2(\sqrt{A_s}/f)} \;\Rightarrow \frac{f}{\sqrt{A_s}} = \frac{2n_{cycle}R}{w} \tag{8.56}$$

To illustrate the use of equations (8.50) and (8.56) we shall work out a specific example. *Example*: Let us consider applying the equations in a first order design of a thermal camera that should allow the *recognition of a tank* with a geometrical profile of 2.5x2.5 meters at a distance of 3 km with a frame rate of 30 frames/sec. We have the following subsystem and scene parameters:
- The total field of view to be covered is 1 degree.
- A single sensor element with dimensions 50x50 μm.
- An operational wavelength bandwidth 8 - 12 μm.
- A spectral *exitance* $\partial M(\lambda,T)/\partial T = \pi\partial L(\lambda,T)/\partial T$, at a wavelength of 10 μm, of $5{\cdot}10^{-5}$ W cm^{-2} μm^{-1} K^{-1} computed from the Planck equation.
- A waveband averaged D^* of $2{\cdot}10^{10}$ cm Hz$^{1/2}$W^{-1}.
- The requirement for a NETD value of 0.1 K.

Question: What are the required optical properties in terms of front-e a certain subjective elemnd diameter of the aperture lens and the camera focal length?
By employing equation (8.56) we can calculate the focal length of the system, with the requirement of 4 resolved cycles on target, as 48 cm and α_{pix} =104 μrad, corresponding to a spatial frequency ξ =4808 cycles/radian. This ξ-value results in a resolving power of 10 *line pairs/mm* in the image plane of the thermal camera. A one degree field of view contains 168x168 104 μrad sensor elements (pixels). With the imposed frame rate of 30 frames/sec this amounts to an exposure (dwell) time of $\approx$ 1.2 μsec per sensor element. Applying the Nyquist criterion $\Delta\nu = 1/(2\Delta t)$ this yields a bandwidth $\Delta\nu$ = 417 kHz. Substituting this bandwidth and the above numerical values for the various parameters in equation (8.50), we arrive at a focal ratio $(f/\#)$ =0.88, hence the required lens aperture equals 54.5 cm. It is obvious that such a large lens diameter is impracticle, so we need to deploy a focal plane array of sensor elements. This greatly increases the dwell time and reduces drastically the frequency bandwidth to 15 Hz. As a result the focal ratio can be raised to $(f/\#)$ = 11.4, yielding a camera aperture diameter of 4.2 cm.

The diffraction limit of a circular aperture with this diameter is λ/D = 238 μrad at 10 μm, implying an MTF that accommodates a maximum spatial frequency ξ up to 2100 cyles/radian leading to 4.4 *line pairs/mm* resolving power in the image plane. This is less than half the spatial frequency required to meet the objective of a 0.1 K NETD. In conclusion, a design result from a radiometric computation solely based on a predefined

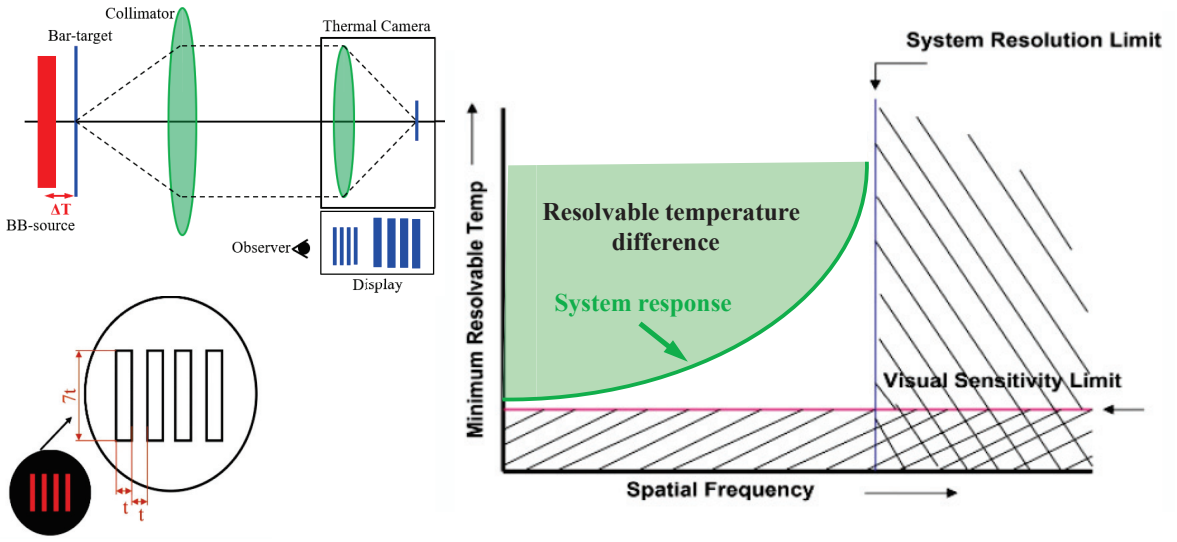


Figure 8.44: *Recognition test with a 4-bar target(4 stops, 4 open spaces) template. The target is back-lit by a uniform extended-blackbody source (red glow). The target is placed at the focus of a high-quality collimator and thus appears to be at infinity, simulating the infinite conjugate case for which the thermal imager is typically designed.*

NETD target value, only incorporating geometrical optics, is clearly inadequate for design purposes since it does not account for the optics OTF. As already introduced above this led to the Minimum Resolvable Temperature Difference (MRTD) as a more relevant design parameter. Minimum resolvable temperature difference (MRTD) is a measure for assessing the performance of infrared cameras, and is inversely proportional to the modulation transfer function. Typically, an operator is asked to assess the minimum temperature difference at which a 4-bar target can be resolved. The MRTD is therefore an *observed thermal sensitivity* as a function of spatial frequency, it includes the effect of spatial integration, temporal integration, noise and image resolution. It also includes, however, interpretation of the 4-bar-image by a human observer during the assessment, which introduces a subjective element in the result. This effect can be partly averaged out by including a number of different observers in the measurement sequence. Smaller bar targets will require larger temperature differences to be visible because of the fall-off in the MTF at higher spatial frequencies. This is schematically shown in figure (8.44) for a four-bar target width w at range R with spatial frequency $\xi_{rec} = 4/(w/R)$. At low spatial frequencies, the MRTD approaches the thermal sensitivity limit (horizontal asymptote), the high spatial frequency cut-off (vertical asymptote) is the frequency value where the system MTF goes to zero, i.e. the system resolution limit, so the MRTD goes to infinity. The green area covers the region of resolvable temperature differences which is bounded by the MRTD(ξ) curve.
Constructing an analytical model for the MRTD is complicated owing to the many parameters that potentially play a role. Measurements indicate that the MRTD apparently does not depend on the size A_s of the sensor element, in contrast to the NETD. On the other hand it does depend on the spatial frequency of the bar target ξ_b and on the system MTF(ξ) of the thermal imager and the noise level embodied by the NETD. We

can therefore write the following proportionality relation for the MRTD as a function of spatial frequency ξ_b which includes the main variables of interest:

$$\begin{aligned} MRTD(\xi_b) &\propto \frac{\xi_b\sqrt{\alpha_{hpix}\cdot\alpha_{vpix}}}{MTF(\xi_b)}NETD \\ MRTD(\xi_b) &\propto \frac{\xi_b\sqrt{\alpha_{hpix}\cdot\alpha_{vpix}}}{MTF(\xi_b)}\left[\frac{(f/\#)^2\sqrt{\Delta\nu}}{D^*(\partial L/\partial T)\sqrt{A_s}}\right] \end{aligned} \tag{8.57}$$

The solid angle term for a rectangular single bar slit/stop of the 4-bar target, ($\alpha_{hpix}\cdot\alpha_{vpix}$), neutralizes the pixel area A_s-term in the NETD denominator, rendering MRTD insensitive for the area of the sensor element, but keeping its dependence on all the other NETD parameters. Moreover the MRTD rises potentially faster with frequency than the decline of the MTF(ξ_b) in the denominator of expression (8.57) due to the frequency term ξ_b in the numerator. The detection and recognition ranges can be predicted from an MRTD-curve for a given target size,target temperature and background temperature, given an atmospheric attenuation. Water vapor (H_2O) and carbon dioxide (CO_2) in the air are the primary causes of absorption of heat radiation that, consequently, loses some of its intensity before reaching the thermal imager. However absorption usually affects the background more than the objects in the foreground. The water vapor content of the air affects the image quality even in sunny and clear weather. During winter, if all other weather conditions are the same, the water vapor content of the air is lower than during the summer period, so this results in an image with better quality during winter than it would have been on a summer day. Moreover, the thermal radiation the object emits will also be scattered when it hits particles. The loss of radiation is directly related to the size and concentration of these particles: droplets or crystals (fine dust) that constitute polluting, condensing, or precipitating conditions, such as smog, fog, rain, or snow. In dense fog, the water droplets are bigger due to piling (or accretion), thus scattering thermal radiation more than light fog. Fog scatters thermal radiation to a larger extent than both smog and haze because of the greater size and concentration of its water droplets.
It is of practical interest to measure the MRTD without the need for a human observer, the subjective link in the measurement chain. Such an *objective* MRTD measurement would be faster and more cost effective than measurements that include an observer each time. In that context, equation (8.57) can be written in a generalized way:

$$MRTD(\xi_b) = F(\xi_b)\frac{NETD}{MTF(\xi_b)}, \tag{8.58}$$

where the constant of proportionality and any spatial-frequency-dependent terms (including the observer) is taken up in the function $F(\xi_b)$. To characterize the average effects of an observer, for a given display and viewing geometry, an MRTD-curve is measured for a representative sample of the system under test. Along with the MRTD data, the NETD and the MTF are also recorded for this system. From these data, the function $F(\xi_b)$ can be determined and subsequent tests of similar systems can be performed without an observer.