1 Introduction

Semiconductor technology forms the basis of the modern electronics industry. The current drive toward miniaturisation is fuelled by the demand for faster switching speeds, lower power requirements and higher integration. As devices become smaller, they become more sensitive to the effects of minor imperfections in the crystal lattice. These imperfections, called *defects*, are formed either during the growth of the semiconductor crystal or during subsequent processing steps such as metallization, ion-implantation, annealing and etching, and can affect the performance of devices.

In most cases, defects in semiconductors are detrimental to device performance, reducing the mobility of carriers and acting as trapping and recombination centres that reduce free carrier lifetime. However, there are a number of applications in which defects are used to enhance device performance, e.g. electron irradiation induced defects in Si power diodes increase the switching speed of the diodes by acting as recombination centres at the *p-n* junction that reduce minority carrier lifetime (Baliga, 1996). In order to predict the influence that a defect would have on device performance (be it detrimental or beneficial) it is essential that the properties of the defect are known. Intimate knowledge of the defect's characteristics would allow the use of defect engineering to improve the device characteristics by eliminating detrimental defects and in some cases even introducing defects that enhance device performance. Peaker (1993) discusses some examples of defect engineering.

One of the most important techniques used to determine the electrical characteristics of a defect is known as Deep Level Transient Spectroscopy (DLTS). During a DLTS measurement, the emission of carriers from a deep-level trap is investigated as a function of temperature. From this data, it is possible to draw an Arrhenius plot, from which one obtains the apparent capture cross-section σ_a and the position of the energy level in the band gap, E_T. These two quantities are collectively known as the defect's *signature*, which can help to identify the defect. A DLTS measurement can also reveal other properties of the defect such as its depth distribution and introduction rates, as well as metastability and annealing properties.

The standard DLTS technique makes use of a lock-in amplifier (LIA) to analyse the signal from the sample. However, with modern technology it has become possible to digitise the signal and analyse it numerically. Digital data analysis has a number of advantages over the older analogue method: Firstly, because no repetitive signal is required, it is possible to analyse much slower transients that would otherwise take unreasonably long in an LIA system. Secondly, in principle only a single transient is required to do a complete analysis over the whole frequency range – this allows for much faster data acquisition and the recording of “single shot” events as are observed with metastable defects. Thirdly, the digital data analysis technique allows much more advanced data analysis to be performed. This technique would allow the resolution of the DLTS system, which is inherently limited with an LIA-based DLTS system, to be improved using inverse Laplace transforms and other deconvolution techniques (see, for instance Dobaczewski, 1994 or Istratov, 1997).

In this thesis, the design and construction of a digital, isothermal DLTS system is discussed. The system has been evaluated firstly by measuring a number of “standard” defects that have been well described in the literature and secondly by using it to analyse a number of defect phenomena that would not be observable by a standard LIA-based system. The results of these measurements are discussed and some publications, in which these techniques have been used, are included.

In Chapters 2 and 3, the general theory of defects and deep-level transient spectroscopy is discussed. The discussion is not intended to be a complete study, but rather to highlight a number of topics, which will be referred to at a later stage and to define nomenclature and notation that will be used in the rest of the thesis. In Chapter 4 the design and characterisation of the digital DLTS system is discussed, while Chapter 5 describes the general experimental procedures that were followed. The experimental results are presented in Chapters 6 to 9. The experimental chapters consist of an introduction describing the basic experiment and theoretical background, followed by more detailed experimental procedure and results. Where applicable, a copy of published papers containing the discussed results and conclusions, have been included at the end of the chapter.

Chapter 10 contains general conclusions and suggestions for further research.

2 Some concepts in semiconductor physics

The properties of semiconductors are adequately described in a number of textbooks. For this reason only a short description of the aspects of semiconductor physics that are relevant to this study is given. The main aim is to familiarise the reader with the terminology and the notation used in this thesis. For a more complete introduction to the properties of semiconductors, the reader is referred to textbooks e.g. Ridly (1988) , Sze (1981), Smith (1978), and Henisch (1989).

2.1 Metal-semiconductor junctions

A number of early researchers have noted that the current flowing through a metal–semiconductor junction depends on the polarity of the applied voltage. This effect was researched further and later used in point contact rectifiers. Currently, metal-semiconductor junctions are important because they are used in devices as well as tools in the analysis of physical parameters of semiconductors. For this reason, metal-semiconductor junctions have been studied extensively.

A number of models have been suggested to explain how these junctions operate. In this chapter, the model proposed by Schottky (1942) will be discussed in more detail. This model describes an ideal case, where the metal and the semiconductor are in intimate contact, without the presence of any interfacial layer or interface states. The Bardeen model describes a more general case where the effects of an interfacial layer and interface states are taken into account (Bardeen, 1947 and Rhoderick, 1988).

2.1.1 The Schottky model

When a metal contact is evaporated onto the surface of a semiconductor, a potential barrier is formed at the metal–semiconductor interface. Here, only the case for an *n*-type semiconductor will be considered. The formation of a Schottky barrier on *p*-type material occurs similarly.

2.1.2 The ideal case

Figure 2.1 graphically illustrates the formation of a Schottky barrier. Part (a) illustrates the metal and the semiconductor in their isolated, electrically neutral states. Here χ_s is the electron affinity of the semiconductor. (The electron affinity of a substance is the energy released when an electron is added to the material – i.e. in this case the difference between the vacuum level and the conduction band edge.) ϕ_m and ϕ_s are the work functions of the metal and the semiconductor respectively. (The work function of a material is the energy required to remove an electron from the material to the vacuum level – i.e. the difference between the vacuum level and the Fermi level.) Here we only consider the case where the work function of the metal is greater than that of the semiconductor, which, in practice, is the most important case. This relationship between the two work functions causes the Fermi level of the metal to be lower than that of the semiconductor, and leads to the formation of a contact with rectifying properties.

Now, if the metal and the semiconductor were connected by means of a thin wire, electrons would flow from the semiconductor to the metal due to the difference in work function. Because of this flow of electrons, a positive charge builds up on the surface of the semiconductor, while a negative charge builds up on the surface of the metal, causing an electric field in the gap between the metal and the semiconductor. This electric field opposes the flow of electrons. The equilibrium condition is reached when the Fermi levels of the two materials coincide. This implies that the potential difference between the metal and the bulk of the semiconductor is equal to the difference in their Fermi levels.

The negative charge that builds up on the surface of the metal is caused by extra electrons that are accommodated within the Thomas-Fermi screening distance of about 0.5 Å, i.e. within the first atomic layer. In the semiconductor, the positive charge is caused by the removal of electrons. However, the only electrons close to the Fermi level that can be removed are those in the conduction band, which are provided by the ionised donor atoms. Thus, the positive charge in the semiconductor is provided solely by the uncompensated donor atoms, left after electrons have flowed from the conduction band.

The concentration of these donor atoms is much lower than the concentration of electrons in the metal. This means that electrons are depleted from the conduction band up to an appreciable depth, *w*. For carrier densities of 10^{16} cm^{-3} the thickness of this so-called depletion layer, is generally in the order of a micron. Because the charge in the depletion region is distributed over a finite distance, the potential changes slowly over the depletion region, and causes the bands to bend upwards as shown in Figure 2.1(b).

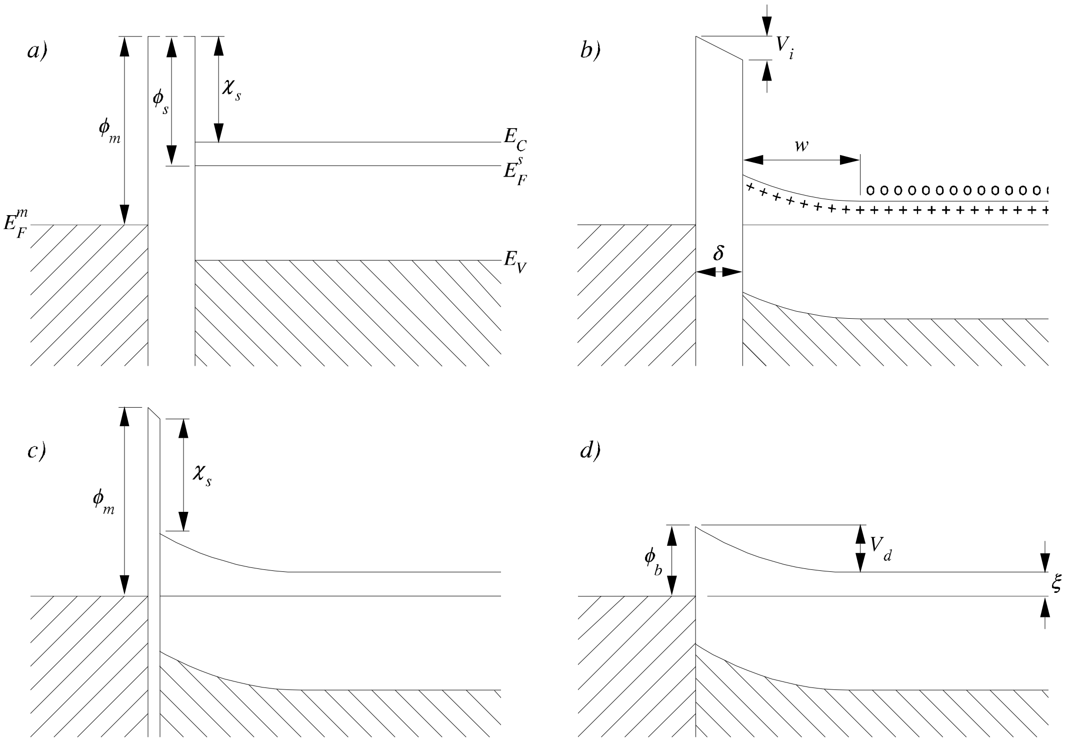

Figure 2.1 *The forming of a Schottky barrier. a) The metal and the semiconductor in their isolated states, b) electrically connected, c) separated by a narrow gap, d) in perfect contact. (After Rhoderick, 1988.)*

The difference between the electrostatic potentials at the surfaces of the metal and the semiconductor is given by $V_\text{i} = \delta E_\text{i}$, where δ is the distance between the metal and the semiconductor, and E_i is the electric field in the gap. As the gap, δ, is decreased, the electric field stays finite [Figure 2.1(c)], and causes V_i to tend towards zero as the gap disappears. When ideal contact is made, the barrier due to the vacuum disappears completely [Figure 2.1(d)], and the only barrier seen by electrons, is that caused by the bending of the bands in the semiconductor. The height of this barrier measured relative to the Fermi level is given by

$$\phi_\text{b} = \phi_\text{m} - \chi_\text{s} \,. \tag{2.1}$$

The height of the barrier relative to the position of the conduction band in the neutral region of the semiconductor is called the diffusion potential (also called the built-in voltage), and is indicated by V_d. From Figure 2.1(d) it is clear that, under zero bias conditions

$$V_\text{d} = \phi_\text{b} - \xi \,, \tag{2.2}$$

where ξ is the energy difference between the Fermi level and the conduction band in the neutral semiconductor. From charge neutrality, it can be shown that (Sze, 1981)

$$\xi = kT \ln \frac{N_\text{C}}{N_\text{D}} \,, \tag{2.3}$$

where N_C is the density of states in the conduction band of the semiconductor and N_D is the doping density.

In practice, it is difficult to fabricate Schottky diodes by conventional vacuum deposition without a thin insulating layer of oxide about 10 to 20 Å thick forming on the surface of the semiconductor. This layer is often referred to as the interfacial layer. A practical Schottky diode is therefore better represented by Figure 2.1(c). The interfacial layer is usually very thin, so that electrons can easily tunnel through it, causing this case to be almost indistinguishable from the ideal case illustrated in Figure 2.1(d). Furthermore, the potential drop V_i is so small that Equation (2.1) remains a reasonable approximation.

In this discussion, we have made a number of assumptions that are not always valid. For example, we have neglected the effect of interface states. For a more complete discussion on other aspects influencing the barrier height, the reader is referred to Cowley (1965) and Rhoderick (1988).

2.1.3 Behaviour of the barrier under forward and reverse bias

When a bias is applied across the barrier, the relationship between the Fermi levels in the semiconductor and the metal is changed. Under zero bias conditions, the electrons from both sides of the junction “see” the same barrier height relative to their Fermi energy. As a result, there is no net flow of electrons over the barrier in the one or the other direction.

However, if a forward bias is applied (i.e. a positive potential is applied to the metal), the position of the Fermi level in the semiconductor is raised relative to that of the metal. This decreases the amount of band bending, causing electrons in the semiconductor to “see” a lower barrier than those in the metal. This causes a net flow of electrons from the semiconductor to the metal. As the forward bias is increased, the barrier presented to the electrons decreases, causing an increase in the current flowing over the junction.

Under reverse bias, the Fermi level of the semiconductor is lowered relative to that of the metal, causing more band bending and an increase in the barrier seen by electrons in the semiconductor. This also increases the width of the depletion region. The barrier experienced by electrons from the metal, however, stays constant. Thus, the current from the metal to the semiconductor stays constant, as the current from the semiconductor to the metal decreases. This causes the current under reverse bias to tend toward a limit as the reverse bias is increased. This continues until the electric field in the depletion region is large enough to cause dielectric breakdown of the semiconductor, leading to a large current flowing across the barrier. This could cause irreversible damage to the device. For a detailed discussion on current transport mechanisms, see Rhoderick (1988) or Sze (1981).

From the discussion in the previous section, it follows that the relationship between V_d and the bias applied to the diode, V_D, can be written as

$$V_\mathrm{d} = \phi_\mathrm{b} - \xi - V_\mathrm{D} \,. \tag{2.4}$$

2.1.4 Calculation of the electric field in a Schottky barrier

The shape of the band edge profiles can be calculated by solving Poisson's equation, subject to certain boundary conditions. The first boundary condition is obtained from the barrier height, while the second is that there is no electric field in the bulk of the semiconductor. By choosing $x = 0$ at the interface, the boundary conditions can be written as $V(0) = V_{\mathrm{d}}$ and $E(\infty) = 0$. These relationships serve as boundary conditions for the solution of Poisson's equation in the semiconductor, which can be written in the one-dimensional case as

$$\frac{d^2V}{dx^2} = \frac{1}{\varepsilon_{\mathrm{s}}}\rho(x), \tag{2.5}$$

where $\rho(x)$ is the total charge density in the semiconductor at a depth x and ε_{s} is the permittivity of the semiconductor. In general, contributions from the valence band, the conduction band, ionised donors and acceptors and deep levels in the band gap should be taken into account. This however leads to a very complicated equation that can only be solved numerically. The equation can be simplified by making use of the depletion approximation.

According to the abrupt or depletion approximation, it is assumed that it is possible to divide the semiconductor into two regions: the depletion region, directly below the metal, which is devoid of free carriers, and the bulk of the semiconductor, which is electrically neutral, and in which no electric field exists. In the depletion region, where there are no electrons in the conduction band, the charge density, $\rho(x)$ is qN_{D}. If the width of the depletion region is *w*, the charge density in the semiconductor can be written as

$$\rho(x) = \begin{cases} qN_{\mathrm{D}} & \text{if } x \le w \\ 0 & \text{if } x > w \end{cases}. \tag{2.6}$$

By integrating Equation (2.5) twice, and applying the boundary conditions, the width of the depletion region can be obtained as

$$w = \sqrt{\frac{2\varepsilon_{\mathrm{s}}V_{\mathrm{d}}}{qN_{\mathrm{D}}}}, \tag{2.7}$$

while the electric field and potential in the depletion region are given by

$$E(x) = -\frac{qN_{\mathrm{D}}(w-x)}{\varepsilon_{\mathrm{s}}} \tag{2.8}$$

and

$$V(x) = -\frac{qN_D}{2\varepsilon_{\mathrm{s}}}(w-x)^2. \tag{2.9}$$

Figure 2.2 shows a graph of $\rho(x)$, $E(x)$ and $V(x)$ for a typical Schottky barrier.

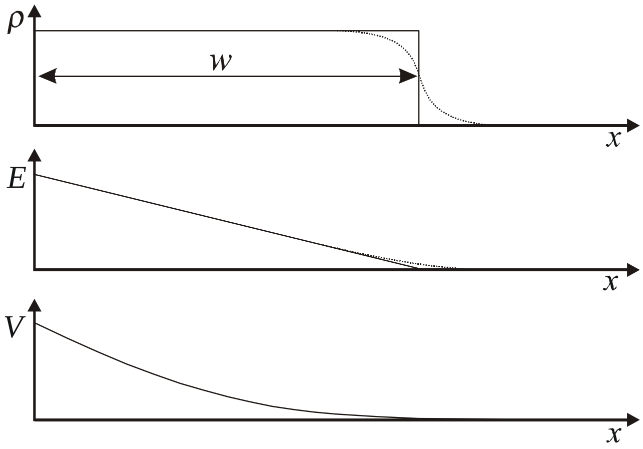

***Figure 2.2** Graphs of the charge density ρ, electric field E, and electrostatic potential V in the depletion region of a Schottky diode. The solid line indicates the results according to the depletion approximation, while the dashed line indicates the effect of a non-zero Debye length.*

The charge per unit area in the depletion region is now

$$Q_\mathrm{d} = qwN_\mathrm{D} = \sqrt{2q\varepsilon_\mathrm{s} N_\mathrm{D} V_\mathrm{d}}\ , \tag{2.10}$$

from which it follows that the (differential) capacitance per unit area of the diode is

$$C = \frac{\partial Q_\mathrm{d}}{\partial V_\mathrm{d}} = \sqrt{\frac{\varepsilon_\mathrm{s} q N_\mathrm{D}}{2V_\mathrm{d}}} = \frac{\varepsilon_\mathrm{s}}{w}\ . \tag{2.11}$$

In practice, this capacitance is measured by superimposing a small oscillating voltage ΔV_osc on the applied bias and measuring the current due to the voltage. The capacitance is then calculated from the relation $C = \Delta Q / \Delta V$. Since the properties of a Schottky diode are highly non-linear, it is important that ΔV_osc is small compared to the applied bias, so as not to influence the measurement.

It should be noted that the depletion region approximation is exactly that – an approximation. In reality, the transition between the depletion region and the bulk is not abrupt, but consists of a gradual transition region, as shown by the dashed line in Figure 2.2. In this transition region the potential decreases approximately exponentially with a decay constant L_D, called the Debye length Jackson (1975):

$$L_D = \sqrt{\frac{kT\varepsilon_\mathrm{s}}{q^2 n}}\ , \tag{2.12}$$

where n is the carrier concentration in the semiconductor. The effects of a non-zero Debye length are particularly important when the depletion region edge is used to measure depth profiles, where the depth resolution of the technique is limited by the Debye length.

2.2 Defects and deep levels in semiconductors

The band structure of semiconductors is generally calculated for an ideal crystal (i.e. one that contains an infinite number of repetitions of the unit cell without any deviations) at 0 K. In reality, all materials contain some chemical impurities or lattice defects. Such an interruption in the lattice periodicity of a crystal is called a defect. These defects can be classified as point defects in which the perturbation of the lattice remains localised (e.g. vacancies, interstitials, substitutions) or extended defects (dislocations, surfaces, grain boundaries, voids, cavities). Since small aggregates of several point defects (e.g. divacancies, vacancy–donor complexes etc.) also cause only a local perturbation of the lattice, they are generally considered as point defects as well.

Since the periodicity of the crystal lattice is an important factor determining the band structure of the solid, one can expect that a defect will in some way influence the eigenvalues of the Schrödinger equation. Many of these defects cause bound states to appear in the forbidden energy gaps of the perfect crystal. These bound states have wave functions that decrease exponentially away from the defect. Except for numerical techniques, there is no general theory for calculating the energy levels of these bound states, but a number of approximations apply in different limiting cases.

The states in the forbidden band gap can broadly be classified into “shallow” and “deep” states. The shallow states are close to the edges of the forbidden band and can be described using the effective mass theory for which the bound-state equation reduces to a hydrogenic Schrödinger equation. On the other hand, deep level defects, which lie closer to the middle of the band gap, are better described by the tight binding approximation. According to this approximation, the eigenstates of the bound electrons are expressed as a linear combination of the free-atom eigenstates. This method is frequently referred to as the linear combinations of atomic orbitals (LCAO) method. The main advantage of the LCAO method is that it is relatively simple but none the less gives a qualitatively correct description of many experimental observations. For a detailed description of these and other methods, see Lannoo (1981), Jaros (1982) or Ridley (1988).

2.2.1 Emission and capture of carriers from defects

Defect states in the band gap can influence the properties of a semiconductor in a number of ways. Except for behaving as donors or acceptors, defect states may also influence the mobility of charge carriers by scattering, and cause various features in the optical absorption and emission spectrum of the semiconductor. The most important electrical effect of deep levels in the band gap of a semiconductor is the emission and capture of charge carriers. These processes cause various transient effects and cause defects to act as recombination and trapping centres, influencing the carrier lifetimes in semiconductors.

The kinetics of emission and capture of carriers from defect levels has been discussed extensively in the literature, see Shockley (1952), Hall (1952) and Bourgoin (1983). In this section, the case of a single level with two charge states in a non-degenerate semiconductor will be discussed, similarly to the approach followed by Bourgoin (1983).

Consider a defect with two charge states *S* and *B*, where in the *S* state the defect contains one more electron than in the *B* state. The notation e_n, e_p, k_n and k_p will be used for the probabilities for emission (e) and capture (k) of electrons (n) and holes (p).

If one assumes that there is no barrier that the electron has to overcome during the capture process, the probability per unit time, k_n, that a defect in state *B* captures an electron from the conduction band can be written as

$$k_\mathrm{n} = c_\mathrm{n} n = \sigma_\mathrm{n} v_\mathrm{th,\,n} n\,, \tag{2.13}$$

where c_n is the electron capture coefficient of the defect, σ_n is the electron capture cross section of the defect, $v_\mathrm{th,\,n}$ is the thermal velocity of the electrons in the conduction band and n is the concentration of the electrons in the conduction band.

A similar equation can be written for holes:

$$k_\mathrm{p} = c_\mathrm{p} p = \sigma_\mathrm{p} v_\mathrm{th,\,p} p\,, \tag{2.14}$$

where p is the concentration of holes in the valence band.

If the carrier concentration is much less than the density of states in the conduction and valence bands, the number of empty states in the conduction and valence band is approximately independent of the carrier concentration. Therefore, it may be assumed that the emission rates e_n and e_p are independent of the carrier concentration. Now, if s and b are the concentration of defects in state *S* and *B* respectively, the rates for emission and capture of holes and electrons are given by:

$$\begin{array}{ll} k_\mathrm{n} s = c_\mathrm{n} n s & \text{electron capture} \\ e_\mathrm{n} b & \text{electron emission} \\ k_\mathrm{p} b = c_\mathrm{p} p b & \text{hole capture} \\ e_\mathrm{p} s & \text{hole emission} \end{array} \tag{2.15}$$

At thermal equilibrium, the capture rates for each species should be equal to its emission rate, i.e. $c_\mathrm{n} n^0 s^0 = e_\mathrm{n} b^0$ and $c_\mathrm{p} p^0 b^0 = e_\mathrm{p} s^0$, where the superscript (^{0}) indicates values at thermal equilibrium. It is now possible to solve for e_n and e_p, giving

$$e_\mathrm{n} = c_\mathrm{n} n^0 \frac{s^0}{b^0} = \sigma_\mathrm{n} v_\mathrm{th,\,n} n^0 \frac{s^0}{b^0} \tag{2.16}$$

and

$$e_\mathrm{p} = c_\mathrm{p} p^0 \frac{b^0}{s^0} = \sigma_\mathrm{p} v_\mathrm{th,\,p} p^0 \frac{b^0}{s^0}\,. \tag{2.17}$$

At thermal equilibrium, the ratio between the concentrations of the two charge states of the defect is

$$\frac{s^0}{b^0} = \gamma \exp\left(\frac{E_T - E_F}{kT}\right), \tag{2.18}$$

where γ is a degeneracy factor equal to $Z(\mathrm{S})/Z(\mathrm{B})$. Furthermore, at thermal equilibrium the carrier densities are Bourgoin (1983)

$$n^0 = N_\mathrm{C} \exp\left(\frac{E_\mathrm{F} - E_\mathrm{C}}{kT}\right) \tag{2.19}$$

and

$$p^0 = N_\mathrm{V} \exp\left(\frac{E_\mathrm{V} - E_\mathrm{F}}{kT}\right). \tag{2.20}$$

Substituting Equation (2.18) into Equations (2.16) and (2.17), and using Equations (2.19) and (2.20), it follows that

$$e_\mathrm{n} = \sigma_\mathrm{n} v_\mathrm{th,\,n} \gamma N_\mathrm{C} \exp\left(-\frac{E_\mathrm{C} - E_\mathrm{T}}{kT}\right) \tag{2.21}$$

and

$$e_\mathrm{p} = \frac{\sigma_\mathrm{p} v_\mathrm{th,\,p} N_\mathrm{V}}{\gamma} \exp\left(-\frac{E_\mathrm{T} - E_\mathrm{V}}{kT}\right). \tag{2.22}$$

Assuming a Boltzmann distribution, the thermal velocity of electrons in the conduction band can be written in terms of their effective mass m_e^* (see, for instance Mandl, 1988)

$$v_\mathrm{th,\,n} = \left(\frac{8kT}{\pi m_\mathrm{e}^*}\right)^{1/2} \tag{2.23}$$

and the density of states in the conduction band N_c is

$$N_c = \frac{1}{\sqrt{2}}\left(\frac{m_e^* kT}{\pi \hbar^2}\right)^{3/2}. \tag{2.24}$$

By substituting Equation (2.23) and (2.24) into Equation (2.21), the emission rate can be written as

$$e_\mathrm{n} = \frac{2}{\pi^2 \hbar^3} \frac{\sigma_\mathrm{n} m_\mathrm{e}^{*2} k^2 T^2}{g} \exp\left(-\frac{E_\mathrm{C} - E_\mathrm{T}}{kT}\right). \tag{2.25}$$

Here, in agreement with the more common notation (see, for instance Miller, 1977), the degeneracy factor γ used by Bourgoin (1983) has been replaced by $1/g$, where g is the degeneracy of the defect level.

If it is assumed that the capture cross-section of the defect is independent of temperature, it follows that an Arrhenius plot of $\ln(e_\mathrm{n}/T^2)$ as a function of $1/T$ should yield a linear relationship from which the defect's energy E_T and capture cross-section σ_n may be calculated. These two values are frequently referred to as the defect's signature. The defect signature is one of the important parameters used to identify a defect during electrical measurements.

Note, however, that the capture cross-section calculated from an Arrhenius plot is subject to a number of assumptions, for example, that the capture cross-section of the defect is not temperature dependent. Therefore, the capture cross-section calculated from the Arrhenius plot is frequently referred to as the *apparent* capture cross-section and indicated by $\sigma_{\mathrm{n,a}}$ in order to distinguish it from the capture cross-section determined by more direct means. For the same reason the energy level of the defect calculated from the Arrhenius plot, might differ from values obtained under different conditions or by other techniques. The differences between values obtained using different techniques, might give important information about the nature of the defect involved.

2.2.2 Defect occupation as a function of time

Many experiments measure the emission and capture of defects after the equilibrium concentration has been disturbed in some way. Consider a semiconductor containing a single deep level with a concentration N_{T}. The concentration of occupied defects is N. Due to the emission and capture of carriers, the concentration of occupied defect will change according to

$$\frac{\mathrm{d}N}{\mathrm{d}t} = (c_{\mathrm{n}} + e_{\mathrm{p}})(N_{\mathrm{T}} - N) - (c_{\mathrm{p}} + e_{\mathrm{n}})N \; . \qquad (2.26)$$

The general solution to this differential equation is

$$N(t) = N(\infty) + [N(0) - N(\infty)]\exp[-(c_{\mathrm{n}} + e_{\mathrm{p}} + c_{\mathrm{p}} + e_{\mathrm{n}})t] \, , \qquad (2.27)$$

where $N(\infty)$ is the equilibrium concentration of the occupied defect for which $\mathrm{d}N/\mathrm{d}t = 0$ and is given by

$$N(\infty) = \frac{c_{\mathrm{n}} + e_{\mathrm{p}}}{c_{\mathrm{n}} + e_{\mathrm{p}} + c_{\mathrm{p}} + e_{\mathrm{n}}} \, . \qquad (2.28)$$

It is often helpful to divide defects into two classes, namely minority and majority carrier traps. Majority carrier traps are defects for which the thermal emission rate for majority carriers e_{maj} is much greater than the thermal emission rate for minority carriers e_{min}. For a minority carrier trap, the opposite is true, i.e. $e_{\mathrm{min}} \gg e_{\mathrm{maj}}$. The terms *electron* trap and *hole* trap are frequently used to distinguish defects for which $e_{\mathrm{n}} \gg e_{\mathrm{p}}$ and $e_{\mathrm{p}} \gg e_{\mathrm{n}}$, respectively. Clearly, an electron trap in an *n*-type semiconductor is a majority carrier trap, while an electron trap in a *p*-type semiconductor would be a minority carrier trap.

In thermal capture and emission experiments, usually only one of the emission rates dominates the kinetics, so that Equation (2.27) may be simplified considerably. For example, for an electron trap in the depletion region, the emission rate e_{n} dominates all the other emission and capture rates, so that defect concentration as a function of time is given by

$$N(t) = N_{\mathrm{T}} \exp(-e_{\mathrm{n}}t) \; . \qquad (2.29)$$

2.2.3 Field dependence of the emission rate

In Section 2.1.4, it was mentioned that an electric field exists in the depletion region of a Schottky diode. This field may be quite large, almost up to the dielectric breakdown field of the semiconductor ($\sim 10^7$ V/m). If a defect is placed in the depletion region of a Schottky diode, it will also be subject to the field, which will distort the shape of its potential well. This distortion of the potential well may enhance the emission probability of a carrier trapped in the well. The extent of this enhancement depends, amongst others, on the shape and dimensions of the potential well. A few enhancement mechanisms will be discussed briefly in the following sections.

2.2.3.1 The Poole-Frenkel effect

The simplest mechanism according to which emission of an electron from a potential well may be enhanced is the Poole-Frenkel effect. When an external field is applied to an electron trapped in a potential well, the electron is subjected to the sum of both fields. This causes the shape of the potential well to be distorted, thus raising the barrier on the one side of the defect and lowering it on the other (see Figure 2.3).

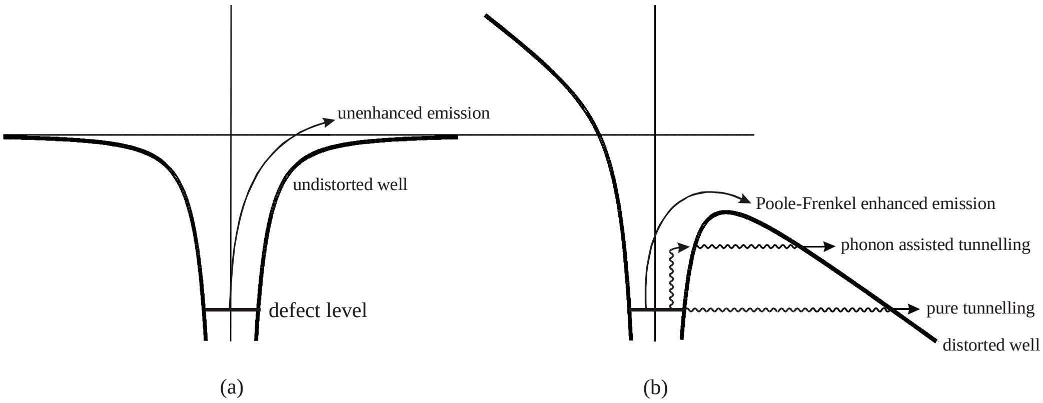

***Figure 2.3** The distortion of a coulombic well by an electric field. (a) The undistorted (zero-field) coulombic well and (b) the coulombic well in a constant, external electric field. A number of emission mechanisms are indicated schematically.*

The original theory, as developed by Frenkel (1938) deals with the one-dimensional case only. According to the one-dimensional model, the ionisation energy of a coulombic well placed in an electric field *F*, is lowered by

$$\Delta E_T = \sqrt{\frac{qF}{\pi\varepsilon}} . \tag{2.30}$$

When substituted in Equation (2.21), this implies that the emission rate of the defect is now given by

$$e(F) = e(0)\exp\left(\frac{1}{kT}\sqrt{\frac{qF}{\pi\varepsilon}}\right), \tag{2.31}$$

where $e(0)$ is the emission rate at zero electric field, k is Boltzmann's constant, T is the absolute temperature.

However, the one-dimensional model over-estimates the emission enhancement. Hartke (1968) developed the theory for a three-dimensional well, according to which the emission rate from a three-dimensional coulombic well can be described by:

$$e(F) = e(0)\left\{\left(\frac{kT}{\beta\sqrt{F}}\right)^2\left[1 + \left(\frac{\beta\sqrt{F}}{kT} - 1\right)\exp\left(\frac{\beta\sqrt{F}}{kT}\right)\right] + \frac{1}{2}\right\} \tag{2.32}$$

where

$$\beta = \left(\frac{q^3}{\pi\varepsilon}\right)^{\frac{1}{2}}. \tag{2.33}$$

The emission enhancement due to the Poole-Frenkel effect is frequently used by experimentalists to estimate the range of the defect potential. A longer range potential (e.g. a coulombic well) would show a far stronger Poole-Frenkel enhancement than a shorter range potential.

The characteristic dependence of the emission rate (e) on electric field (F) in the case of the one-dimensional Poole-Frenkel effect for a coulombic well, namely that log e is proportional to $F^{1/2}$, has been used as experimental evidence to distinguish between donor and acceptor defects. The linearity of this dependence is characteristic of a charge leaving a centre of opposite sign. In n-type material this would imply a donor type defect, whereas, in p-type material this would imply an acceptor type defect (Bourgoin, 1983).

2.2.3.2 Phonon-assisted tunnelling

The field-enhanced emission due to the Poole-Frenkel effect is generally relatively small. Some defects, however, show a much stronger field enhanced emission. This strong field enhanced emission may be explained by tunnelling mechanisms. The two mechanisms discussed here are "pure" tunnelling and phonon-assisted tunnelling, with the pure tunnelling mechanism being predominant in the high field regions (>10^8 V m^{-1}).

The phonon assisted tunnelling mechanism is observed in defects with a significant electron–lattice coupling. Due to this coupling, a trapped electron can occupy a set of stationary quasi levels separated by $\hbar\omega$, with $\hbar\omega$ being the phonon energy. Elastic tunnelling can then occur from any of these quasi deep levels to the conduction band. The coupling constant or Huang-Rhys factor S (Makram-Ebeid, 1980) is represented by

$$S = \frac{\Delta E}{\hbar\omega}, \tag{2.34}$$

where ΔE is the vibrational energy loss. The field emission rate due to phonon assisted tunnelling emission as derived by Pons (1979) is given by

$$e_f = \sum_p \Pi_p \Gamma(\Delta_p).(1 - f_{1,p}) \, . \tag{2.35}$$

The $(1 - f_{1,p})$ factor in the equation is the Fermi-Dirac probability of finding an empty conduction band state, $\Gamma(\Delta_p)$ is the tunnelling emission probability for an electron at a quasi level *p* with an energy Δ_p above the ground state and Π_p is the probability of finding the electron at quasi level *p*.

The tunnelling probability $\Gamma(\Delta_p)$ was calculated by Korol (1977) for an electron trapped in a delta function potential well as

$$\Gamma(\Delta) = \gamma \frac{\Delta}{qK} e^K \, , \tag{2.36}$$

where Δ is the energy position of the deep level below the conduction band and *K* is the WKB attenuation of the wave function across the potential barrier separating the trapping site from the free conduction band states. The pre-exponential factor γ is equal to $q/3\hbar$. Assuming a uniform field *F* and a triangular barrier, *K* is given by

$$K = \frac{4}{3} \frac{\sqrt{2m^*}}{\hbar F} \Delta^{3/2} \, , \tag{2.37}$$

where m^* is the electron effective mass.

The probability Π_p of finding an electron at a given quasi level $E_c - \Delta_p$, where $p = 0, \quad \pm 1, \quad \pm 2, \ldots$, may be calculated from

$$\Pi_p = (1 - e^{\hbar\omega/kT}) \sum_{n=0}^{\infty} e^{n\hbar\omega/kT} J_p^2 \left(2\sqrt{S\left(n + \tfrac{1}{2}\right)} \right) , \tag{2.38}$$

where J_p is a Bessel function of the first kind and *n* the integer number of phonons. This model is based on the assumption that the phonons have a single, well-defined angular frequency (ω).

In practice, the theoretical model can be fitted to experimental emission rate vs. electric field data recorded at different temperatures in order to obtain experimental values for the parameters *S* and γ.

3

DLTS: Deep level transient spectroscopy

Deep levels in semiconductors influence both the electrical and the optical properties of the material. There are a number of optical techniques for the characterisation of deep level defects in semiconductors. However, one of the main shortcomings of these techniques is that they cannot measure or predict the electrical properties. Since most semiconductor applications rely on the electrical properties of the semiconductor, it is important to know the electrical properties of a deep level defect. Furthermore, many of the processes that occur in deep levels that influence device performance are nonradiative, and cannot be observed by optical techniques. Deep level transient spectroscopy (DLTS) is one of the most versatile techniques used to determine the electrical properties of defects.

3.1 The DLTS technique

As described by Lang (1974), the DLTS technique uses a fast, sensitive capacitance meter to measure the capacitance of a reverse-biased Schottky, MOS or *p-n* junction. According to Equation (2.11), the capacitance of a reverse-biased diode can be related to the width of the depletion region, which in turn depends on the charge in the depletion region, due to dopants as well as deep levels. The DLTS technique measures the change in the capacitance of the junction due to the emission of carriers by defects in the depletion region, as described by Equation (2.29). By processing the capacitance signal with a weighting function, the emission rate of the defect in the depletion region is obtained.

Consider the case of a Schottky contact on an *n*-type semiconductor, as shown in Figure 3.1(A). The semiconductor contains a low concentration of a defect that causes a deep electron trap with energy E_{T}. In the figure the bulk of the semiconductor containing free carriers is indicated by the shaded area, while the depletion region is left unshaded. Filled and open circles indicate filled and empty traps respectively. For simplicity, we assume that, initially, all the traps in the depletion region are empty, while all the traps in the bulk of the semiconductor are filled.

At the start of the DLTS cycle, a smaller reverse bias (or even a forward bias) pulse is applied across the diode. [Figure 3.1(B)]. This bias pulse reduces the width of the depletion region, increasing the capacitance of the Schottky diode drastically. The reduction in the width of the depletion region fills the traps up to a distance of approximately the depletion width below the surface of the semiconductor.

After the filling pulse, the reverse bias is returned to its quiescent level [Figure 3.1(C)]. This increase in reverse bias increases the width of the depletion region again. However, since some of the deep level traps in the depletion region are now filled, the charge density in the depletion region is less than it was in (A), therefore the depletion region is slightly wider and the capacitance slightly lower than was the case in (A).

Since the filled traps in the depletion region are above the Fermi level, they now emit carriers by means of thermal processes, as described in Section 2.2. This causes the charge density in the depletion region to increase, reducing its width and increasing the capacitance of the junction.

If it is assumed that $N_{\mathrm{T}} \ll N_{\mathrm{D}}$, the depletion region width will not change significantly during the emission of carriers. Under these circumstances, it is reasonable to assume that the emission of carriers from the depletion region may be described by an exponential decay, as in Equation (2.29). The capacitance of the Schottky diode is then also described by an exponential decay function

$$C(t) = C_{\infty} + \Delta C e^{-\lambda t}\,, \tag{3.1}$$

where λ is the decay rate and C_{∞} is the steady state capacitance of the diode.

If it is assumed that all the defects from the depletion region edge to the interface are filled by the filling pulse and subsequently emptied, the defect concentration may be calculated from the amplitude of the exponential decay function by applying Equation (2.11)

$$N_{\mathrm{T}} \approx 2N_{\mathrm{D}}\frac{\Delta C}{C}\,. \tag{3.2}$$

It is possible to obtain an activation energy and a capture cross-section associated with the emission of the carriers from the defect by measuring the decay time constant as a function of temperature, as described in Section 2.2.1.

In the above explanation, it has been assumed that the defect level is empty in the depletion region and full in the bulk. Since the defect level typically lies much deeper in the band than the dopant level, the defect level intersects the Fermi level a distance λ shallower than the depletion region edge, as shown in Figure 3.2.

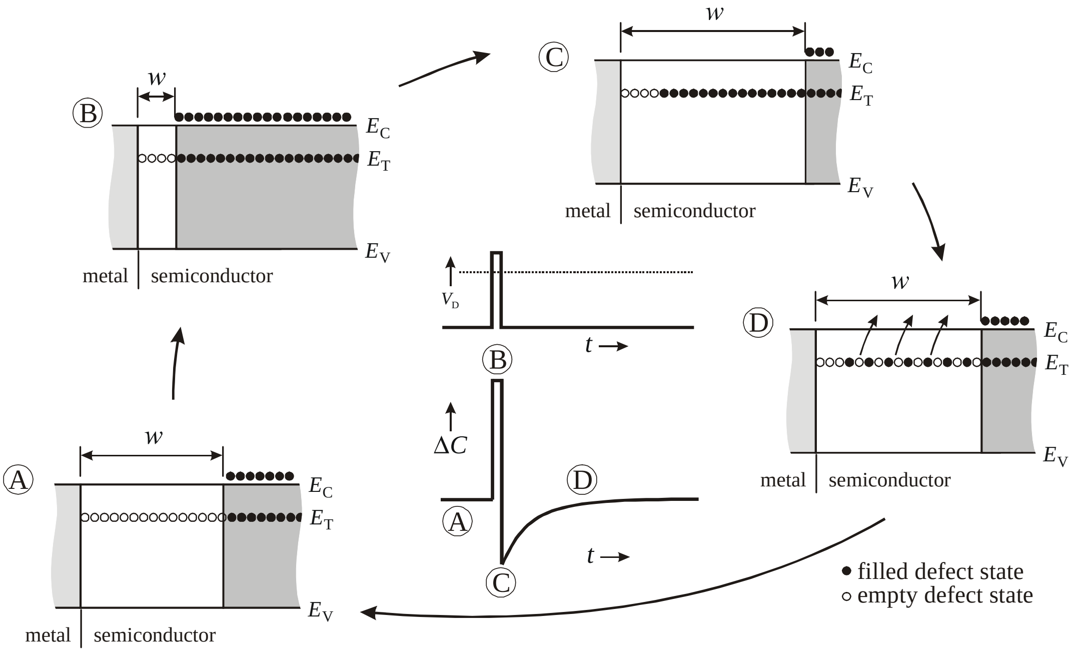

Figure 3.1 *A schematic illustration describing the origin of the DLTS transient. (A): Quiescent state, (B): Filling pulse; (C) Reverse bias; (D) Exponential decay as carriers are emitted. The graphs in the middle show the applied bias* V_D *and the change in the capacitance of the diode* ΔC *as a function of time. (After Miller, 1977.)*

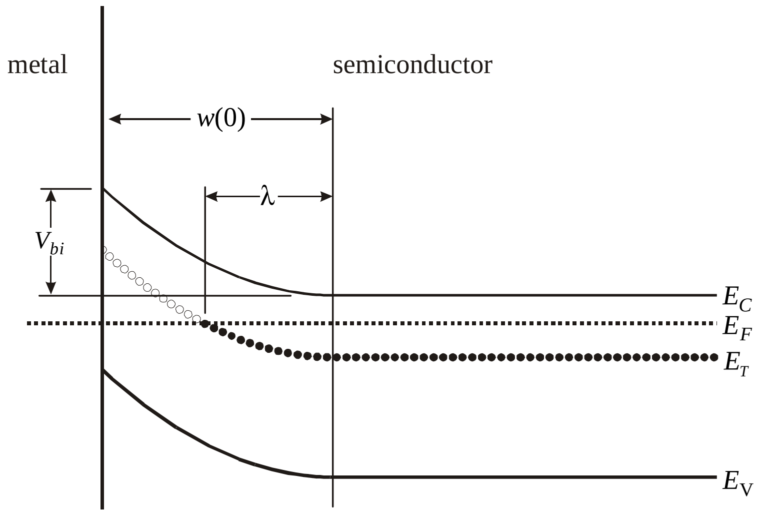

Figure 3.2 *An energy band diagram showing the relationship between the depletion region edge and the position where the defect level intersects the Fermi level.*

If a constant dopant concentration N_D is assumed, then the distance λ is independent of the width of the depletion region, and is given by

$$\lambda = \sqrt{\frac{2\varepsilon_\mathrm{s}(E_\mathrm{C} - E_\mathrm{T})}{qN_\mathrm{D}}} \tag{3.3}$$

The result is that the DLTS measurement does not probe the region at the depletion region edge, but a region a distance λ shallower, as shown in Figure 3.3. This effect has to be taken into account when the DLTS technique is used to determine the depth profile of defects or the electric field experienced by the defects is calculated. Furthermore, during the transient, charges are removed a distance λ from the depletion region edge, therefore Equation (3.2) is not strictly valid, and a more careful analysis needs to be performed in order to obtain quantitatively correct values. DLTS depth profiling and the precautions that need to be taken are discussed in detail by Zohta (1982).

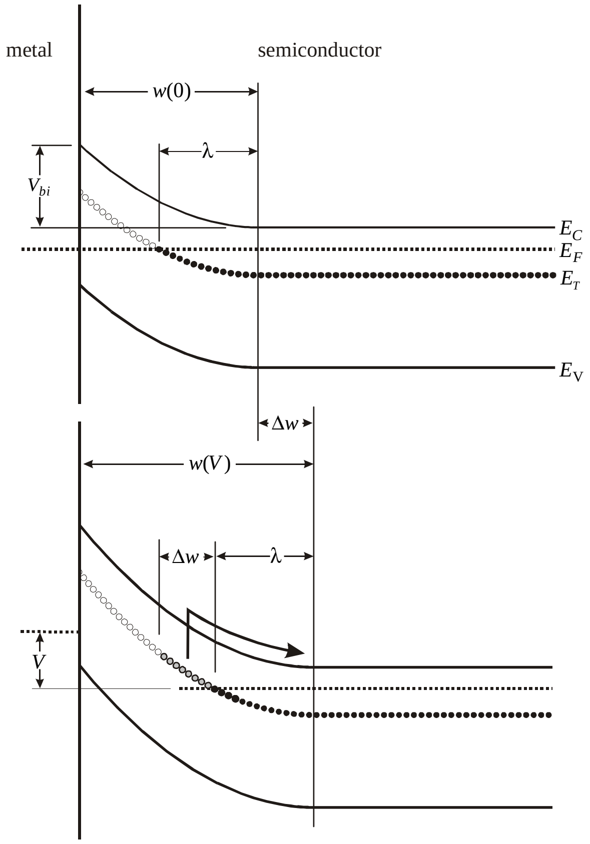

***Figure 3.3** Filling (top) and subsequent emission (bottom) of electrons from a deep level in the depletion region, assuming a constant Fermi level. In the top diagram, the filling of the defect during the filling pulse is shown. Note that due to band bending and the depth of the defect level, the defect level is filled to a depth λ shallower than the depletion region edge. After the filling pulse, the width of the depletion region increases by an amount Δw. Carriers are now emitted from a band of defects with a width Δw a distance λ shallower than the depletion region edge.*

3.2 Analysis of the DLTS transient

In the idealised case, the DLTS capacitance transient obtained from a sample with a single defect level is an exponential decay function of the form

$$C(t) = A\mathrm{e}^{-\lambda t} + C_{\infty}\ , \tag{3.4}$$

where A is the amplitude, λ is the decay rate and C_{∞} is the capacitance of the junction at equilibrium. Many physical processes are described by such an exponential decay process, and in principle determining the values of A, λ and C_0 is reasonably straightforward.

However, it frequently occurs that there is more than one defect level in the semiconductor and that these defects have closely spaced decay constants, or even a continuous band of decay constants. In this case, the capacitance transient can be described as the sum of a number of exponential decay functions. Many techniques have been developed for the deconvolution of such a multi-exponential function. However, there are significant problems associated with all of the techniques, and the analysis generally becomes unreliable in the presence of noise.

3.2.1 Analogue techniques

The original method used to analyse the DLTS transient was a double boxcar proposed by Lang (1974). According to this method, the DLTS signal is obtained by subtracting the capacitance measured at time t_2 from the capacitance at time t_1 (both times measured relative to the filling pulse). Qualitatively the process may be described as follows: Assume that the sample is at a low temperature and therefore there is a slow transient. Because the capacitance does not change much, the DLTS signal $S = C(t_1) - C(t_2)$ is very low [Figure 3.4(a)(i)]. As the temperature is increased, the decay rate of the transient increases causing a greater change in the capacitance between times t_1 and t_2. Therefore the DLTS signal increases as the temperature is increased [Figure 3.4(a)(ii – v)]. This increase in the DLTS signal continues until the transient decays so fast that most of the decay occurs before t_1. A further increase in temperature will now decrease the DLTS signal [Figure 3.4(a)(vi – x)]. When the DLTS signal is plotted as a function of the temperature, as in Figure 3.4(b), a peak is observed. The expression for the time constant at which the maximum in the DLTS signal is observed is easily derived, and depends on the values of t_1 and t_2 :

$$\lambda_{\max} = \frac{\ln(t_2 / t_1)}{t_2 - t_1}\ . \tag{3.5}$$

Because the capacitance transient is very small, it is important to minimise the effect of noise on the measurements. For this reason, the capacitance measurements are usually averaged over a number of transients and, instead of taking a single point, the average capacitance values for an interval around t_1 and t_2 are taken.

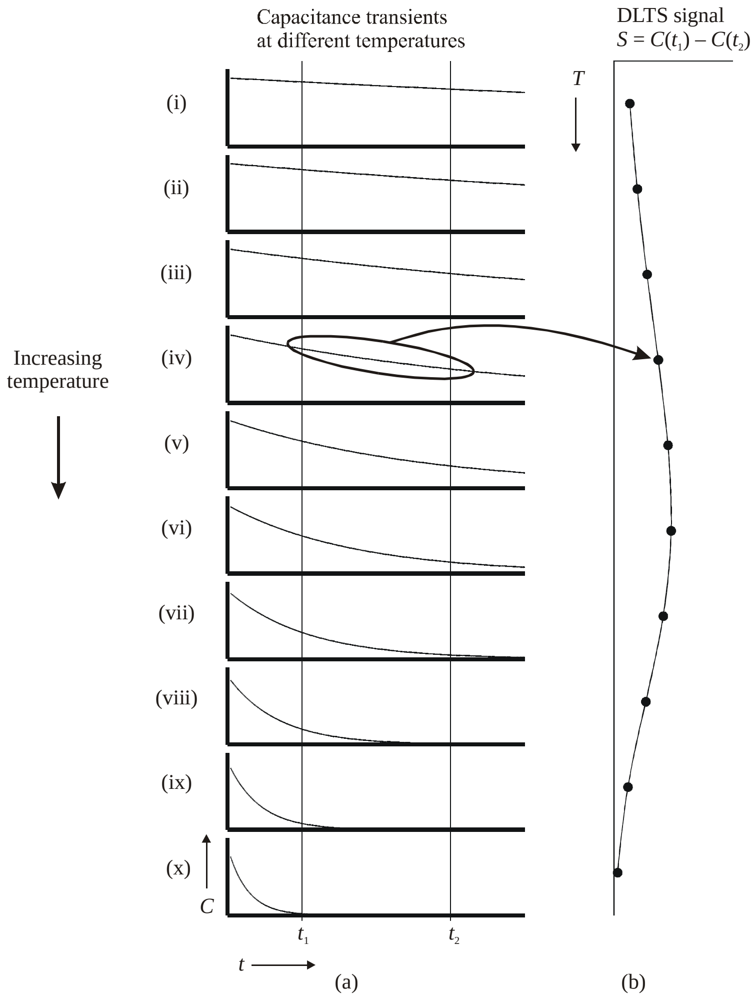

Figure 3.4 *(a) The change in the shape of a DLTS transient with increasing temperature and (b) the DLTS signal obtained from the transients plotted as a function of sample temperature. (After Lang, 1974)*

In most modern analogue DLTS systems, a lock-in amplifier is used to analyse the DLTS transient. In this arrangement, the transient is convoluted with a sine wave of a fixed frequency according to

$$S(\tau) = \frac{1}{\tau}\int_0^\tau C(t)\sin\left(\frac{2\pi t}{\tau}\right)dt\ . \qquad (3.6)$$

The result of this convolution is referred to as the DLTS signal and is plotted as a function of the sample temperature. The result obtained is similar to that obtained from the double boxcar method, except that since the lock-in amplifier uses more of the signal the lock-in amplifier method is less sensitive to noise than the double boxcar method. It can be shown that, for an exponential transient with a sine wave weighting function, the DLTS signal reaches a maximum when $\lambda = 1/(0.423\tau)$. It is also possible to use other weighting functions, with varying degrees of success.

As an alternative to the above method, it is possible to keep the temperature constant and rather change the frequency of the LIA. This technique is referred to as isothermal DLTS or frequency scanned DLTS (Henry, 1985). Isothermal DLTS is usually implemented using digital signal processing, where a fast analogue to digital converter (ADC) connected to a computer is used to capture and analyse the data.

Both the boxcar and lock-in amplifier techniques can be successfully implemented using a fast ADC and a computer. Using a computer to analyse the transient has the further advantage that analysis at a number of lock-in amplifier frequencies can be done simultaneously, thereby saving a considerable amount of time. Furthermore, a much wider range of emission rates (especially low emission rates) can be accessed by such systems, allowing defects to be studied over a wider temperature range. However, the main advantage of digital signal processing is that modern techniques to analyse multi-exponential decay functions such as the Gaver-Stehfest method (Istratov, 1997) and various methods for calculating numerical inverse Laplace transforms (e.g. Dobaczewski, 1994) may be used.

3.2.2 Digital processing of DLTS transients

The major shortcoming of the DLTS technique is that, even for emission from a single defect level at a single decay rate, the DLTS peaks as obtained by boxcar and LIA analysis are broad, compared to the typical spacing between defects. This is in stark contrast to optical techniques, which, especially at low temperatures, yield very sharp lines. Although the broad DLTS peaks are not a serious handicap if only a single level is present, it is difficult to accurately determine the emission rate of defects with emission rates spaced less than an order of magnitude apart when both are present in the same spectrum. In fact, analysis by more advanced techniques has shown that DLTS peaks that were previously regarded as a single peak actually consist of a number of discrete peaks that could not be observed due to the limited resolution of LIA DLTS (Dobaczewski, 1992).

Istratov (1998) compares a number of different weighting functions. Some of them reduce peak widths by almost a factor of three compared to widths obtained by LIA DLTS, but at the cost of decreasing the signal to noise ratio by more than an order of magnitude. Using weighting functions based on the Gaver-Stehfest algorithm, Istratov (1997) has shown that it is possible to decrease the peak width by almost a factor 5 (compared to a lock-in amplifier), while still keeping the signal to noise requirements of the input signal realistic.

Another approach to improve the resolution of the DLTS technique is to assume that the sample emits a spectrum of emission rates with a spectral density function $F(\tau)$, so that the capacitance transient can be written as

$$C(t) = \int_0^\infty F(\tau)e^{-t\tau}\,d\tau\,. \tag{3.7}$$

For the case of a single emission rate, $F(\tau)$ is a delta function $\delta(\tau-\tau_0)$, while broader peaks can also be described. The aim is to obtain the $F(\tau)$ corresponding to the measured $C(t)$. In Equation (3.7) $C(t)$ is essentially the Laplace transform of $F(\tau)$, therefore an inverse Laplace transform would be required to obtain $F(\tau)$. The calculation of inverse Laplace transforms is a well-known problem and many techniques exist to do this calculation analytically as well as numerically. However, in the case of

DLTS, a number of complications arise, that make it very hard to find $F(\tau)$. Firstly, $C(t)$ is known only over a limited interval on the positive real axis, therefore the techniques that require the function $C(t)$ to be known on the complex plane cannot be used. It is still possible to calculate $F(\tau)$ under these constraints, however, most of these techniques are extremely sensitive to noise on $C(t)$. The result is that great care has to be taken in developing such an algorithm, and the results obtained should be analysed critically. A successful DLTS system based on such an inverse Laplace transform algorithm has been described by Dobaczewski (1994).

3.3 Differential DLTS

The depth range sampled by the DLTS technique depends on the applied reverse bias and the filling pulse, which respectively determine the maximum and minimum of the depth range that is observed. By recording DLTS transients under different biasing conditions and then subtracting these transients, it is possible to observe defects that lie in a limited depth range only. Such techniques are generally used to measure defect concentration depth profiles, see Lefèvre (1977) or Zohta (1982).

However, this technique is not limited to the measurement of concentration depth profiles. As described by Equation (2.8), the electric field in the depletion region also changes with depth; therefore, defects at different depths beneath the surface experience different electric fields. Thus, by restricting the measurement to only a limited depth range, it is possible to observe the behaviour of the defect under the electric field present at that depth. By applying different bias voltages, it is possible to vary the electric field in the depletion region from approximately zero to almost the breakdown field of the semiconductor. If samples with different doping levels are used, it is possible to study the behaviour of a defect under electric fields that vary by several orders of magnitude, and a variety of phenomena described in Section 2.2.3 may be observed.

4

Planning of the digital DLTS set-up

The change in capacitance of the sample during a DLTS capacitance transient is typically in the order of 1% of the total capacitance. Since it is generally difficult to measure a signal superimposed on such a large background, a lot of noise may be expected in the transient. The problem is not helped by some of the analysis techniques (e.g. inverse Laplace transforms) that are very sensitive to noise and require a high signal to noise ratio of 1000:1 or better. Consequently, great care has to be taken to avoid the effects of both random noise and systematic errors on the system and particular attention needs to be paid to using good quality instruments, correct grounding techniques and adequate screening.

If, despite taking all reasonable precautions, the signal to noise ratio is still too low, random noise may be reduced by taking the average of a large number of transients. This approach is usually followed where very high signal to noise ratios are required, e.g. for Laplace DLTS. However, it is not feasible for very long transients. In this case, it usually suffices to apply a low pass filter to the measured data.

The averaging of a large number of transients will not reduce noise that is periodic with the applied DLTS pulse. This periodic noise may occur either because the transient is recorded at a frequency that is close to that of some source of interference (usually 50 Hz mains noise and harmonics) or that some component of the DLTS system generates periodic noise. The main source of such periodic noise was found to be the pulse generator used to bias the sample.

A further source of error in a DLTS system is temperature measurement and control. Since the emission rate of a defect as well as the capacitance of the diode varies with temperature, it is essential that the sample be kept at constant temperature during the DLTS measurement.

In this chapter, the requirements of some of the components in a DLTS system will be discussed. The system used in this study will be described and the results of some tests performed on the system are presented as well.

4.1 Acquisition of the capacitance signal

The main component distinguishing a digital DLTS system from an analogue LIA-based system is the presence of an analogue to digital converter (ADC) that digitises the analogue signal and sends it to a computer for further analysis. In such a system, a number of trade-offs have to be made. Probably the most important is the speed and the resolution of the ADC.

4.1.1 Random noise

The amplitudes of typical DLTS transients range from about 10^{-1} down to less than 10^{-4} times the average capacitance of the sample, which ranges from 10 to 300 pF. This implies that, in order to be able to observe a transient with amplitude $\Delta C / C$ of 10^{-4}, the system has to measure transients with amplitude as low as 1 fF superimposed on a background of 10 pF. Good results in an LIA type analysis require the noise in the system to be an order of magnitude less than the smallest expected transient, i.e. 0.1 fF. This requirement becomes much more stringent if Laplace analysis is to be performed where a signal to noise ratio of 1000:1 or better is required in order to separate peaks differing by a factor two in emission rate. In order to achieve such a low level of noise, it is important to select a good quality capacitance meter and to ensure that noise from the pulse generator supplying the DLTS bias does not produce extra noise at the output of the capacitance meter.

In most LIA-based DLTS systems, the noise is further reduced by firstly compensating for the background capacitance by means of an “offset capacitor”. This allows the capacitance meter to be used on a more sensitive range, providing a higher signal to noise ratio. Secondly, the output of the lock-in amplifier is filtered with a time constant much longer than the period of the LIA, thereby averaging the signal over a number of transients.

Similar techniques may also be used in a digital DLTS system, except that the analogue filtering of the LIA output is replaced by digitally averaging the output of the capacitance meter over a number of transients. According to the Central Limit Theorem, such averaging should reduce the random noise by a factor of $\sqrt{n}$, where n is the number of averages taken. This averaging technique is very effective for fast transients, but becomes time consuming for longer transients. Here some of the high frequency noise may be removed by applying a smoothing algorithm to the transients.

4.1.2 Response time, sampling rate and resolution

In contrast to a temperature-scanned DLTS system, an isothermal DLTS system keeps the temperature constant while the transient is analysed for decay curves of different time constants. Since the range of peaks that such a system can detect at a specific temperature depends on the range of time constants that can be analysed, it is important to set up the system to measure the widest possible range of decay time constants. In general, there is no lower limit to the speed of an ADC, so only the upper limit will be discussed.

In most capacitance-based DLTS systems, the capacitance meter is the factor limiting the response time. An estimate of the response time can be made by considering a typical capacitance meter operating at a frequency of 1 MHz. These capacitance meters generally need at least ten oscillator cycles to measure the capacitance; therefore, it does not make sense to sample at a rate of more than 100 kHz.

Generally, there is a trade-off between the speed of an ADC and its resolution. It is common to find 8-bit ADCs that operate at frequencies in excess of 100 MHz, however, as soon as a higher resolution is required, the techniques used to speed up these fast ADCs (e.g. flash converters) become prohibitively expensive and can no longer be used.

As mentioned previously, many of the mathematical techniques that are used to analyse a DLTS signal are extremely sensitive to noise and inaccuracies in the DLTS signal, and require a signal to noise ratio of 1000:1. Furthermore, there are frequently noise spikes superimposed on the signal that can be up to ten times the magnitude of the DLTS signal. Therefore, it is rarely possible to use the full range of the ADC for the signal, as some leeway has to be left for the noise spikes. Ideally, the ADC is required to have a resolution of at least 0.01% of full scale (i.e. 4 digits or 14 bits).

4.1.3 Periodic noise

While the averaging technique described in Section 4.1.1 is quite effective in reducing random noise, the technique will not reduce noise that is periodic with the applied DLTS pulse. The main source of such noise was found to be the pulse generator used to apply the DLTS bias to the sample. For example, the HP8115 Pulse Generator produces a glitch of about 0.5 mV roughly halfway through the transient (see Figure 4.2 and associated discussion in Section 4.2.3). Since the period in an LIA-based DLTS system remains constant, such a glitch causes only a shift in the baseline of these systems, for which is easily corrected. However, when frequency-scanned measurements are made or the DLTS signal is analysed numerically, such a glitch halfway through the signal can lead to confusing and misleading artefacts.

4.1.4 Stability

In an isothermal DLTS system, the length of the transient may vary over several orders of magnitude from a couple of milliseconds to tens of kilo seconds. This poses the problem that the instrumentation should be able to respond to fast changes in the order of tens of microseconds while remaining stable and able to record changes taking place over several minutes or even hours.

If the same criterion as for random noise is used, the system has to be stable to a level of approx 1000:1, or $10^5 : 1$ relative to the original capacitance for a typical transient. This implies that the temperature, bias voltage and the drift of the capacitance meter and the ADC should all be tightly controlled.

Equation (2.11) can be used to determine the dependence of the capacitance of a sample on the applied bias. After some manipulation, it follows that

$$\frac{dC}{C} = -\frac{1}{2}\frac{dV_\mathrm{d}}{V_\mathrm{d}} \approx -\frac{1}{2}\frac{dV_\mathrm{r}}{V_\mathrm{r}} . \tag{4.1}$$

From this it follows that in order to measure the capacitance accurately to $\Delta C / C \sim 10^{-5}$, the reverse bias has to remain constant to one part in 2×10^5, or 5 μV in the case of a 1 V reverse bias. A similar procedure can be used to determine the sensitivity of the capacitance to temperature changes, which affect the free carrier density:

$$\frac{dC}{C} = \frac{dN_D}{2N_D} . \tag{4.2}$$

If, as a very rough approximation, one assumes that a 10 K temperature change causes a 10% (linear) change in the carrier concentration, it follows that

$$\frac{dC}{C} \approx 0.05 \frac{dT}{T} . \tag{4.3}$$

This implies that, for a transient at 100 K, the temperature has to be kept constant to 20 mK in order to limit $\Delta C / C$ to less than 10^{-5}. However, at low temperatures close to freeze-out, the capacitance of a sample can change very fast, and even higher temperature stability is required.

Another factor affected by temperature variations is the emission rate. From Equation (2.21), it follows that

$$\frac{de_\mathrm{n}}{e_\mathrm{n}} = -\frac{E_\mathrm{C} - E_\mathrm{T}}{kT}\frac{dT}{T} . \tag{4.4}$$

Clearly the quantity $(E_\mathrm{C} - E_\mathrm{T}) / kT$ depends on the specific defect involved, however, in order for the trap to be observed by DLTS, the depth of the trap $E_\mathrm{C} - E_\mathrm{T}$ has to be significantly greater than kT. On the other hand, if $E_\mathrm{C} - E_\mathrm{T}$ is much greater than kT, emission from the trap will be too slow to be observed. For a typical defect, i.e. the EL2 ($E_\mathrm{T} - E_\mathrm{C} = 0.825$ eV) observed at 300 K, the quantity $(E_\mathrm{C} - E_\mathrm{T}) / kT$ is approximately 30. If $(E_\mathrm{C} - E_\mathrm{T}) / kT$ is increased by 3, the emission rate drops by an order of magnitude. A safe maximum value for this quantity is approximately 50 so that we have

$$\frac{de_\mathrm{n}}{e_\mathrm{n}} \sim -50\frac{dT}{T} . \tag{4.5}$$

I.e. for 10% accuracy in the emission rate, the temperature needs to be constant to 200 mK. Obviously, for techniques such as Laplace DLTS where very narrow peaks are to be resolved, better temperature stability is required.

4.2 Selection and characterisation of the instrumentation

4.2.1 Digitiser

A good compromise between the requirements mentioned above was found in the Agilent 3458A Multimeter. This multimeter uses a multi slope integrating A/D converter that is designed in such a way as to allow for different conversion speeds. For example, at the maximum conversion rate of 100 kHz, the A/D converter provides 16 bits resolution, but at lower conversion rates, the resolution increases up to 24 bits. A further advantage is that the integrating A/D converter takes the average of the input signal over the acquisition time, therefore it automatically smoothes the input signal. This has the advantage that if a signal is sampled at a lower rate (e.g. 10 Hz) the converter does not measure the instantaneous input voltage, but rather takes the average over a 100 ms period, thus smoothing the signal and reducing the noise level. The 3458A is therefore capable of sampling at the required rate with an acceptable resolution, while much higher resolution is available at lower sampling rates.

Furthermore, the short-term stability of the multimeter far exceeds the requirements mentioned in Section 4.1.

4.2.2 Capacitance meter

A Boonton 7200 capacitance meter was used throughout this study. This same model was used in the LIA-based DLTS system. The 7200 has a fast response and a recovery time of less than 50 μs after an overload condition. The capacitance meter also allows the user to set a number of other parameters such as the oscillator signal level and has an internal bias source.

Traditionally, a number of modifications were made to older model Boonton capacitance meters used for DLTS measurements such as the 72B and 72BD. These modifications, described by Wang (1985), Christoforou (1991), and Chappell (1984) speed up the response time of the capacitance meter and reduces the time the capacitance meter requires to recover from an overload. A number of studies were made to compare the response of a modified Boonton 72BD with those of a standard 7200. In all cases, the response time and noise levels of the 7200 were similar or better than that of the modified 72BD, used previously.

In some of the very long period studies where field effect measurements were made, spurious pulses generated by the pulse generator during programming caused unwanted filling of the traps in the sample. It was therefore necessary to protect the sample from the spurious pulses produced by the pulse generator by switching the sample to the internal bias supplied by the capacitance meter. An R-C network was also added to ensure that the sample remained biased while the relay was switching between the two bias sources. The circuit diagram of the modification is shown in Figure 4.1. Note that the R-C network would influence the transition time of the pulses applied to the sample, however since this experiment involved particularly slow transients, and the time constant RC is $\sim 150\ \mu\text{s}$, the effect of the RC circuit could be neglected.

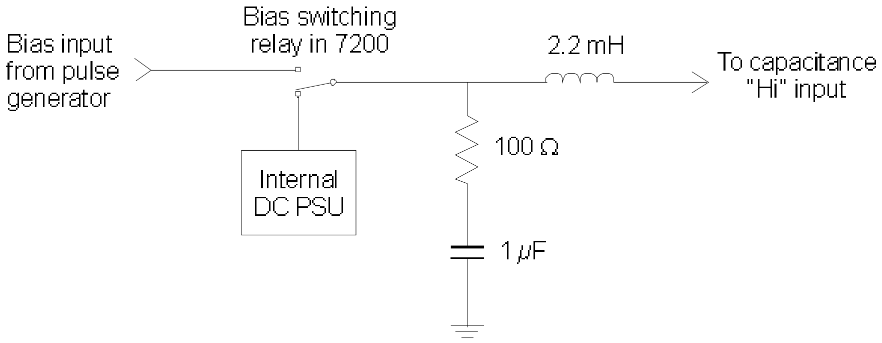

Figure 4.1 *Circuit diagram of the modifications made to the bias circuit of the Boonton 7200 Capacitance Meter.*

4.2.3 Pulse generator

In a DLTS system, the main purpose of the pulse generator is to supply a filling pulse to the sample, followed by a constant quiescent reverse bias during which the capacitance of the sample is observed. In many systems, the pulse generator also supplies the main timing signal and drives accessories such as fast pulse switches and lasers for optical DLTS. A number of pulse generators were available in the department. All were characterised in terms of their output noise and periodic glitches. Due to its programmability and superior performance the Agilent 33120A was selected for most of the work done in this thesis, however where faster pulses were required, an Agilent 8110A was used.

As mentioned earlier, only a small change in voltage is enough to change the capacitance of a Schottky diode by the same magnitude as a typical DLTS transient. Ideally, the voltage across the sample should stay constant to approx. 50 μV, or at least not vary by more than this value in a periodic way. This is especially true for an isothermal DLTS system, where a systematic glitch in the output of a pulse generator, even though its amplitude is much smaller than the random noise, can produce noticeable effects.

Since this requirement on the output of the pulse generator is much more stringent than that required in most applications, it is usually not specified in the specifications of the pulse generator, and it was necessary to measure the output of the pulse generators. During these measurements, the average of 100 pulses was taken in order to reduce random noise and make the periodic features visible. The results of these measurements are shown in Figure 4.2. Clearly, the old HP8115A pulse generator with a 0.5 mV step in the middle of its period would not be suitable. However, even some of the more expensive pulse generators, such as the Agilent 8110A also showed periodic features. The best results were obtained from arbitrary waveform generators, such as the Agilent 33120A and the 33220A, which were relatively free from periodic noise probably because, in contrast to the analogue oscillator, the digital electronics controlling the DAC cannot couple in low frequency periodic noise to the output.

It was finally decided to use the Agilent 33120A pulse generator, which had no noticeable periodic noise after the pulse, and was not too expensive. The UMIST Laplace card also produced a signal that was free from periodic noise, though significantly more random noise. This, however, should not be a problem as the biasing circuit of the capacitance meter easily filters out the high frequency noise.

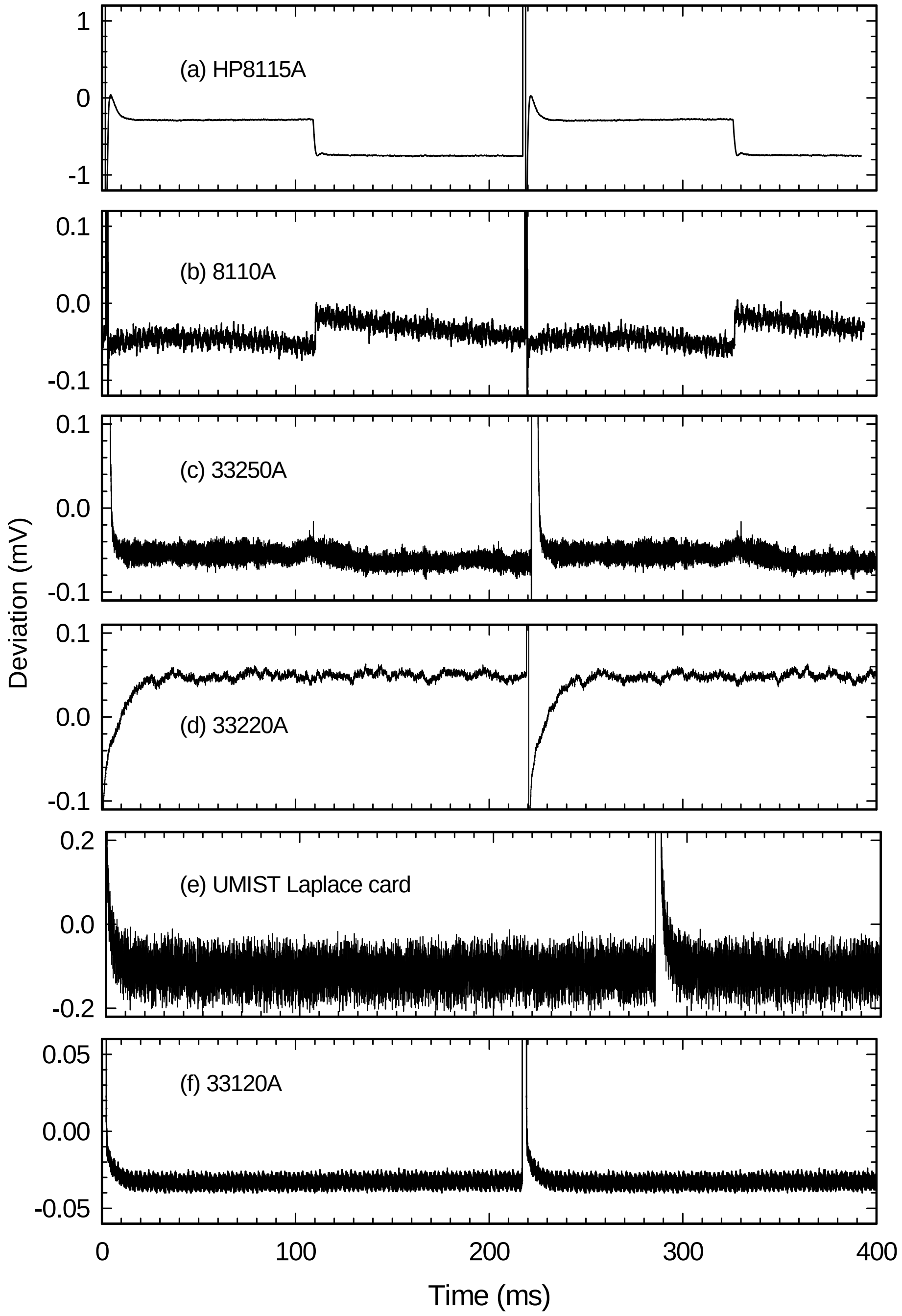

Figure 4.2 *The output obtained from six different pulse generators. In each case, a 1 V pulse was superimposed on a –1 V background. In order to emphasise the periodic noise, the average of 100 pulses was taken. The voltage was measured by means of an HP3458A multimeter on the 1 V scale. Note that the graphs have different vertical scales.*

4.2.4 Rise and fall times

The capacitance meter contains filters that prevent the 1 MHz test signal from travelling along the biasing cables and thus leading to erroneous measurements. However, these filters also act on the external bias applied via the capacitance meter to the diode under test. The positive side effect is that any high frequency noise in the output of the pulse generator is filtered out. However, the high frequency components are also removed from the filling pulse, causing overshoot of the filling pulse seen by the diode under test. While this overshoot is relatively harmless at the trailing edge, at the leading edge of the pulse the overshoot could fill traps shallower than intended. The overshoot can be avoided by smoothing the leading edge of the pulse (Figure 4.3). At the trailing edge, such smoothing is not necessary, and smoothing may in some cases even be detrimental. Therefore, the pulse generator should be able to produce pulses with a slow leading edge and a fast trailing edge. Another reason for the fast trailing edge is that in the current system, the trailing edge was used for synchronisation (see section 4.3.2).

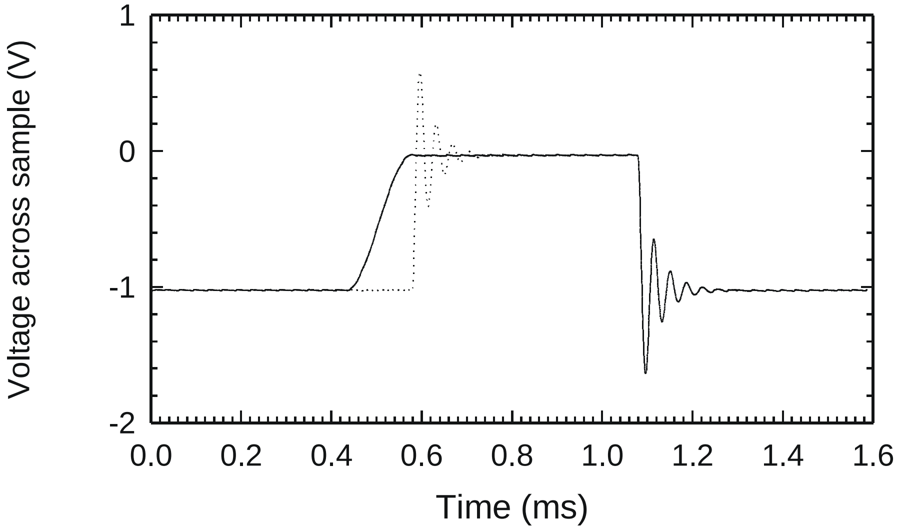

Figure 4.3 *The output of the Boonton 7200 capacitance meter in response to an external bias pulse. The solid line shows the output due to a pulse having a leading edge rise time of 0.5 ms, which eliminated the overshoot indicated by the dotted line, observed when a fast rise time (25 ns) was used. The output was measured using a Nicolet 4570 oscilloscope and the pulse generated using an HP33120A function generator.*

4.2.5 Sample mounting and temperature control

For DLTS measurements, the sample was mounted in an Air Products APD HC-2 cryostat, on a sample holder incorporating a sapphire disc that isolated the sample electrically, but ensured excellent thermal contact with the tip of the cold finger. Myburg (1992) describe the construction and performance of this sample holder (illustrated in Figure 4.4). Contact to the back ohmic contact was established by putting the sample on a piece of indium foil. Both the Schottky contact and the indium foil were connected to the measurement circuit by beryllium copper probes. To minimise damage to the Schottky contact, the tips of the probes were rounded by sanding with very fine sanding paper. In addition, care was taken not to scratch the contact when lowering the probe. To ensure that the ohmic contact made good electrical and thermal contact with the indium foil, an additional probe was used, when necessary, to apply pressure to the sample next to the Schottky contact.

A Lake Shore 340 temperature controller with DT-470 Si diode sensors (calibrated 10 – 500 K) was used for temperature control. Using this system, it was possible to keep the temperature of the sample constant to $\pm 10\ \text{mK}$. In all cases, isothermal measurements were performed only after the temperature stabilised.

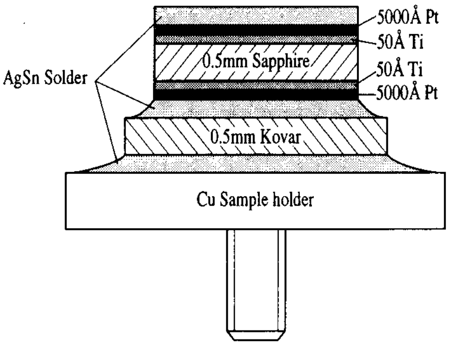

Figure 4.4 *A schematic drawing of the sample holder used during DLTS measurements. (Myburg, 1992)*

4.3 Other hardware

Except for the ready-made instruments, a number of other circuits were required to allow the system to work. For instance, a circuit was required to synchronise the recording of the transient to the DLTS filling pulse and accommodation had to be made for fast pulses, which, due to the filters in the capacitance meter, could not be applied via the capacitance meter.

4.3.1 Fast pulse interface

A number of measurements required the application of short pulses ($< 10\ \mu\text{s}$). These pulses would be severely distorted or even completely filtered out by the filters in the capacitance meter. The solution to this problem is to use reed relays to connect the pulse generator directly to the sample, while at the same time disconnecting the capacitance meter. Since it is critical that the sample is never left unconnected, the timing of the reed relays was set so that the capacitance meter was only disconnected once the pulse generator was connected and there was no more contact bounce from the relay.

Similarly, after the pulse was applied, the pulse generator was only disconnected after the capacitance meter was reconnected.

Great care was taken to select reed relays with short switching times (< 0.1 ms) and minimal contact bounce. A block diagram of the fast pulse interface is shown in Figure 4.5. This circuit allowed pulses as short as 50 ns to pass without significant distortion.

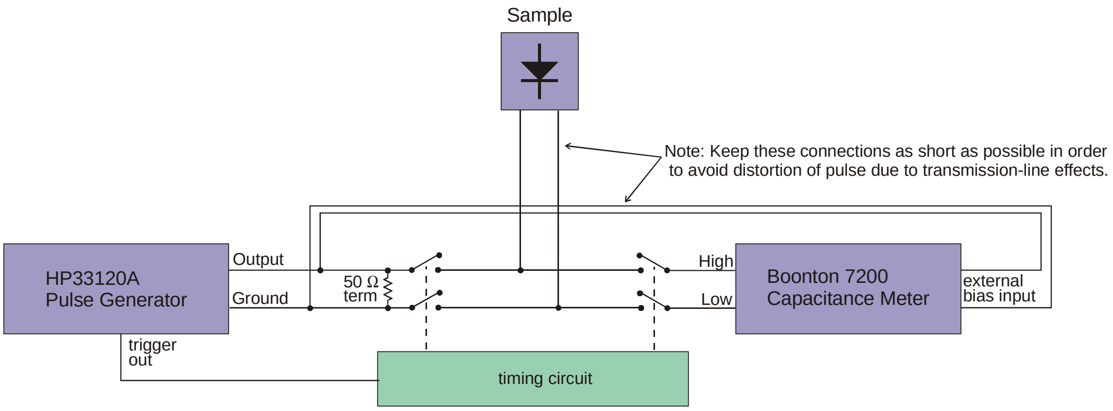
Sample
Note: Keep these connections as short as possible in order to avoid distortion of pulse due to transmission-line effects.
HP33120A Pulse Generator
Output
Ground
trigger out
50 Ω term
High
Low
Boonton 7200 Capacitance Meter
external bias input
timing circuit

Figure 4.5 *A block diagram of the fast pulse interface.*

4.3.2 Synchronisation

It is necessary that the multimeter be in some way triggered at the start of the transient to enable it to start measuring. This trigger has to be accurate to ensure that the same reference point is used for different measurements. Furthermore, when filling pulses of varying lengths are used, the multimeter should always be triggered at a time relative to the trailing edge of the filling pulse. Therefore, it was decided to trigger the multimeter with the trailing edge of the filling pulse. Since the height and offset of the filling pulse can vary, it is not possible to use a level-sensitive trigger. However, since the HP33210A allows for different transition times for the leading and trailing edges of the pulse, the trailing transition of the pulse can be made very fast without affecting the slow rise time required for the leading edge. Consequently, it was decided to trigger on the derivative of the filling pulse. Figure 4.6 shows the circuit that was used. It consists of a voltage follower that acts as a buffer, connected to a differentiator. The output of the differentiator is fed into a voltage comparator (with some hysterisis to avoid oscillations) followed by a monostable timer that eliminated spurious triggering due to oscillations after the initial trigger pulse. The output of this circuit was used to trigger the multimeter as well as an oscilloscope that was used to troubleshoot the set up.

A block diagram of the complete digital DLTS system is shown in Figure 4.7.

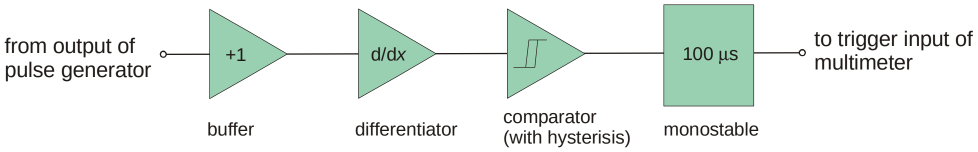
from output of pulse generator
+1
buffer
d/dx
differentiator
comparator (with hysterisis)
100 µs
monostable
to trigger input of multimeter

Figure 4.6 *A block diagram of the trigger circuit that was used for synchronisation.*

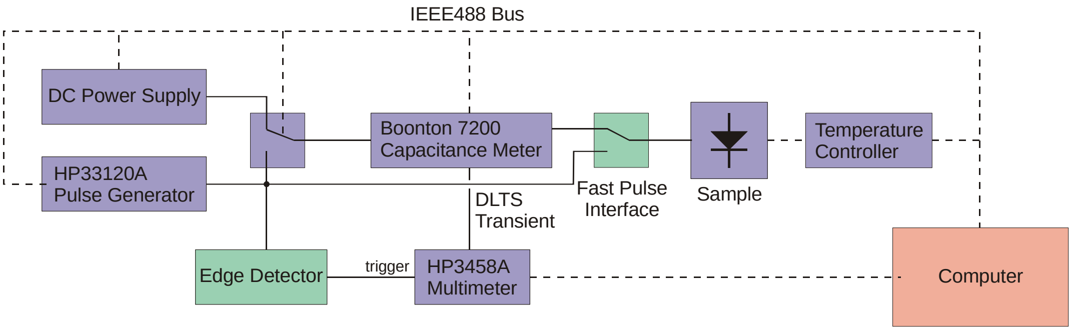

Figure 4.7 *A block diagram of the DLTS system showing the main components.*

4.4 Software

It was decided to transfer data from the multimeter to the computer in real time via the GPIB interface, thereby avoiding the need to use the limited buffer in the 3458A for temporary storage. Furthermore, since the data was transferred during the measurements, no extra time was required for this procedure. The highest transfer rate needed for measurements was 200 kB/s.

In order to achieve this high transfer rate, the software was written in Borland C and run under DOS, where the transfer of data would not be interrupted by other processes. A National Instruments PCI GPIB interface card was used. With current technology, a solution under Windows, using an environment such as LabView is probably possible, and might be preferable to the solution used in this study.

The acquisition software controlled and programmed all instruments that were involved in the measurement. Specifically, the software downloaded the required DLTS pulse to the arbitrary waveform generator, and settings such as sampling rate, resolution and aperture time were set on the multimeter. The acquired signal was averaged during the measurement, and saved to disk. The program allowed up to 64k points per transient to be measured up to the maximum sampling rate of the multimeter, and up to 64k transients to be averaged. Measurements were automated by writing C functions that called subroutines in the main program. These C functions served as confirmation of measurement conditions, while further comments in these C functions were used for explanatory notes and further documentation.

The raw data obtained from the measurement program was smoothed using a linear weighted least squares smoothing algorithm (Jacoby, 2000), in which the data points for the smoothed curve was determined by performing a weighted least squares fit on a window of points surrounding the point in question. The weights of the points were selected to decrease rapidly as points further from the point in question. The advantage of this smoothing technique is that it allows for interpolation so that points in the smoothed data set could be spaced differently to those in the unsmoothed data set. In order to reduce the amount of data involved, the time intervals were spaced logarithmically.

In order to show DLTS spectra over a wide range, a number of transients of differing lengths were recorded. More averages were taken for shorter transients, thereby reducing the noise, while less averages were taken for longer transients, where smoothing could easily reduce noise. The recorded transients were combined to form one long transient during the smoothing procedure. This procedure reduced measuring time, while still producing acceptable DLTS spectra. Care was taken that the filling pulse was of sufficient length to fill all defects, in order to ensure that both the short and the long transients started under identical conditions, so that there was no effect carried over from the previous filling pulse.

Some experiments were made to interlace the recording of the transients. According to this procedure, the longer transient was always preceded by a number of repetitions of the shorter transient, thereby making measurement conditions more similar. However, since this interlacing procedure required more intricate programming, and the more simple procedure of recording the transients of differing lengths separately provided good results, the interlacing was not used.

DLTS spectra were obtained from the smoothed capacitance data by simulating the action of a lock-in amplifier being swept over a frequency range. The DLTS signal was calculated by

$$S(\tau) = \frac{1}{\tau}\int_0^\tau C(t)\sin\left(\frac{2\pi t}{\tau}\right)dt \; . \qquad (3.6)$$

Further manipulation of the signal, such as subtraction and peak detection was performed using the program SigmaPlot.

5 Experimental

5.1 Introduction

The experimental section is organised in four chapters of which this one is the first. Each describes some background theory relevant to the topic. It was decided to rather discuss the aspects of the theory, which are not directly related to DLTS, as part of the experimental section where it is closest to its application.

5.2 Sample preparation

5.2.1 Gallium Arsenide

The same procedures for sample preparation were followed in all experiments. The sample was chemically cleaned before evaporation of the ohmic, as well as the Schottky contacts. The procedures used in both cases were similar, except that cleaning before the evaporation of Schottky diodes, included etching. Myburg (1992) describe the chemical cleaning procedure in detail: It involved degreasing in boiling trichloroethene, followed by rinsing once in isopropanol and then twice in de-ionised water ($\rho > 10^{18}\ \Omega\text{cm}$). Etching was performed, for 30 s, in a solution of 3 parts NH_3 (15 mol dm^{-3}) and 1 part H_2O_2 (30%) in 150 parts de-ionised water at 25°C. Etching was followed by another rinse in de-ionised water, followed by oxide removal in 6 mol dm^{-3} HCl (2 min) and a final rinse in de-ionised water. The samples were dried under a stream of high purity nitrogen, and immediately transferred to the evaporator.

Ohmic contacts were formed by the deposition of Ni (50 Å), AuGe (2000 Å) and Au (2 500 Å) on the n^+ back surfaces of the sample, before the fabrication of the Schottky contacts. The contacts were annealed at 450 °C for 2 min in a quartz tube under a flow of Ar.

Before evaporation of the Schottky contacts, the samples were dipped in 6 mol dm^{-3} HCl for 2 – 3 seconds. Depending on the metal being deposited, either resistive or electron-beam evaporation was used to deposit the contacts through a mechanical mask.

5.2.2 Silicon

Before the deposition of Schottky contacts, the samples were degreased in boiling trichloroethylene followed by rinsing in boiling isopropanol and in de-ionised water. The samples were etched in a 10% HF solution. On the p-type material, Ti Schottky contacts were used.

For ohmic contacts, In/Ga eutectic mixture was used on the unpolished back surfaces.

5.2.3 Gallium Nitride

A similar cleaning procedure was used before deposition of both ohmic and Schottky contacts. The GaN samples were cleaned by boiling them in aqua regia (see Hacke, 1993). After rinsing the samples in de-ionised water, the samples were degreased in boiling trichloroethylene followed by rinsing in boiling isopropanol and in de-ionised water. Hereafter, the samples were dipped in a 50% HCl solution for 10 s, before being transferred to the vacuum system.

Ohmic contacts, consisting of 150 Å/2 200 Å/400 Å/500 Å layers of Ti/Al/Ni/Au (Ruvimov, 1993) were deposited by means of electron beam evaporation and annealed at 500°C for 5 min in Ar. Schottky contacts were deposited by means of either sputter deposition or resistive evaporation through a metal contact mask.

5.3 IV and CV characterisation

A number of artefacts observed during DLTS measurements can be traced to high resistivity and high leakage current (Chen, 1984). In order to evaluate the risk of these effects influencing the DLTS measurements, the quality of the contacts was evaluated by means of IV and CV measurements before DLTS measurements were performed. Since it was found that the quality of the contacts deteriorated after prolonged measurements in the cryostat, the IV and CV measurements were occasionally repeated.

The IV measurements were made by an HP 4148A pA meter/voltage source, capable of measuring currents as low as 10^{-14} A. The sample was screened from light and electrical noise by enclosing it in a light-tight metal box during measurements. The most important characteristics obtained from the IV graph were: the series resistance R_s, the barrier height $\phi_{\mathrm{b(IV)}}$, the ideality factor *n*, and the leakage current at 1 V reverse bias $I(-1)$.

The CV characteristics were measured by an HP 4192A Impedance Analyzer, controlled by a computer via an IEEE interface. The CV measurements were performed with the sample mounted inside the cryostat at the temperature where the isothermal DLTS measurements were to be performed. Measuring conditions (i.e. oscillator level and frequency) were chosen to correspond as closely as possible to those of the DLTS measurements.

The CV characteristics were used to measure the average free carrier concentration of the sample, N_{D}, as well as to check the uniformity of the doping profile. The dissipation factor D (equal to the tangent of the phase angle) was measured at the reverse bias where DLTS measurements were to be performed. This factor was used to determine the degree by which the behaviour of the sample deviates from that of an ideal capacitor.