Capacitance-Voltage Measurement

Capacitance-voltage (CV) measurements are conventionally made using a dedicated test fixture situated within a light excluding enclosure to prevent measurement errors due to extraneous photo-generated currents. Typically, an unpatterned oxide layer is fabricated on a high quality silicon substrate, which is then metallized with a thin film of aluminum. The thin film is patterned to form MOS capacitors, either at the time of deposition by use of a shadow mask or by means of conventional photolithography and chemical etching. Alternatively, MOS capacitors can be fabricated by depositing, heavily doping, and patterning a polysilicon thin film instead of aluminum. The advantage of the use of aluminum is that the deposition and patterning are very convenient; however, heavily doped polysilicon most closely approximates a finished *transistor* structure. Usually, electrical connection to the completed MOS capacitor is made to the topside aluminum or polysilicon contact by means of a thin tungsten probe and to the backside of the wafer by use of a conductive "chuck" which allows a partial vacuum to be drawn under the wafer, *i.e.*, a "vacuum chuck", thus facilitating electrical contact. Although not absolutely necessary for capacitance measurements, it is generally advantageous to remove any pre-existing insulating layers from the back of the wafer. (Indeed, it is usual for a layer of oxide to be grown on the back as well as the front of a wafer during thermal oxidation.) This is easily done by etching in hydrofluoric acid, hydrogen fluoride vapor, or some other method. If backside contact is found to be especially critical, the whole back of the wafer optionally may be metallized. Furthermore, in practice, it is usual for the backside contact, *i.e.*, the silicon substrate, to be held at ground and the frontside contact, *i.e.*, aluminum or polysilicon, to be biased at some potential. In general, it is found that useful information is obtained from CV measurements made for two different conditions, *viz.*, quasistatic and high frequency. (Typically, this data is used to construct a "CV plot" which graphically displays MOS capacitance (or capacitance per unit area) as a function of bias voltage.)

Quasistatic Conditions

As might be expected, quasistatic conditions correspond closely to equilibrium. Conventionally, a quasistatic CV measurement is made by sweeping bias voltage applied to an MOS capacitor so that the surface of the semiconductor changes from inversion to depletion and then to accumulation. During this procedure, displacement current is measured as a function of time. (Obviously, displacement current is just the transient charging current of the capacitor.) Ideally, there is no true conduction current flowing through the oxide layer. Even so, a small amount of "leakage current" invariably flows through any "real" oxide layer. Of course, for quasistatic measurements to be useful the oxide must not be too "leaky", otherwise, conduction current will be added to displacement current and measurements will be inaccurate. (Generally, for a high quality oxide this current is negligible and can be ignored.) Within this context, modern, computer-controlled CV analysis equipment can automatically correct for leakage (if it is not too large or erratic), thus, extending the usefulness of CV measurements to less than perfect oxide layers. Obviously, integration of charging current over time merely results in a measurement of the charge stored in the MOS capacitor; hence, capacitance as a

function of voltage is determined by elementary identification of the product of capacitance and voltage as stored charge, *i.e.*, $CV=Q$.

Clearly, if the surface of the semiconductor is either in accumulation or inversion, a layer of charged mobile carriers is present directly beneath the oxide. (For notational convenience, an italicized *C* denotes capacitance per unit area and a non-italicized C denotes absolute capacitance.) Thus, the measured capacitance, C_{max}, is just the capacitance of the oxide layer alone (equal to $C_{ox}A$ such that, by definition, *A* is the area of the gate, *i.e.*, the frontside contact). In contrast, if the semiconductor is depleted, there is no layer of mobile carriers present underneath the oxide, *i.e.*, at the Si/SiO_2 interface. However, mobile carriers are present underneath the depletion region. Therefore, in depletion, the measured capacitance consists of the series combination of the oxide capacitance and the capacitance of the depletion layer. Of course, this combined capacitance must be less than C_{max}. Accordingly, as the voltage is swept from inversion to accumulation, during depletion the capacitance decreases from C_{max} to a minimum, C_{min}, corresponding to maximum depletion layer width and then rises again as the semiconductor becomes accumulated. (By definition, C_{min} is capacitor area, *A*, times minimum capacitance per unit area, C_{min}.) This essential behavior illustrated in the following figure:

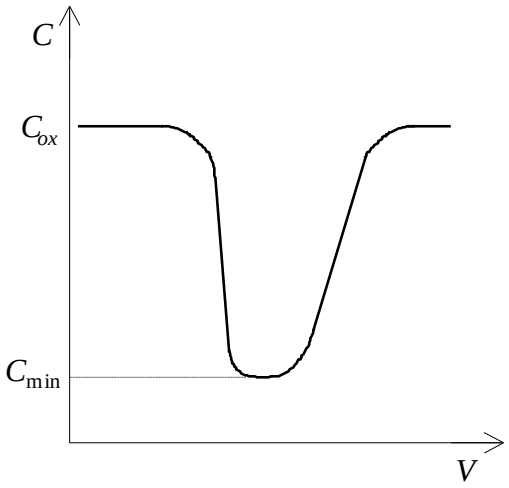

Fig. 41: Idealized quasistatic CV plot (*p*-type substrate; accumulation at left)

Clearly, C_{min} is related to C_{ox} and C_s as follows:

$$\frac{1}{C_{min}} = \frac{1}{C_{ox}} + \frac{1}{C_s} = \frac{x_o}{\varepsilon_{ox}} + \frac{x_d^{max}}{\varepsilon_s}$$

As defined previously, x_o and ε_{ox} are, respectively, thickness and dielectric constant of SiO_2. Likewise, x_d^{max} and ε_s are maximum thickness and dielectric constant of the depletion layer. Of course, electrical charges are still present within the depletion region due to ionized impurity atoms; however, these charges are fixed and do not move in

response to an applied bias (at low temperatures). Consequently, these fixed charges can make no contribution to displacement current and do not result in a contribution to measured capacitance.

High Frequency Conditions

In quasistatic measurements, capacitance is measured directly by integrating charging current. However, CV measurements can also be made by superimposing a small sinusoidally oscillating (AC) signal on the voltage sweep and measuring the corresponding impedance directly as a function of bias voltage. (This requires the use of a high precision impedance meter.) In this case, capacitance measured under conditions of accumulation and depletion can be expected to be the same as observed in quasistatic measurements, *i.e.*, conditions still remain near equilibrium. However, if the frequency of the AC signal is sufficiently high, capacitance measured under a condition of inversion is not the same as in the quasistatic case. The explanation for this is quite simple and is a direct consequence of non-equilibrium behavior of the inversion layer. Physically, any inversion layer must be formed from minority carriers generated in the depletion region and swept to the surface by the electric field. (Of course, minority carriers may be also generated in the bulk and diffuse into the depletion region.) Equilibrium conditions imply that there is sufficient time (by definition) for the inversion layer carrier concentration to respond to any changes in applied field. However, if the material quality of the silicon is good, carrier generation-recombination processes occur very slowly, with a time constant on the order of milliseconds. Therefore, for an applied AC voltage in the megahertz range, the response of the inversion layer is simply too slow to "follow" the signal and similar to ionized dopant impurity atoms, the inversion layer appears fixed with respect to the AC component of the bias. (Of course, the inversion layer does respond to the primary voltage sweep.) This behavior is shown in the following figure:

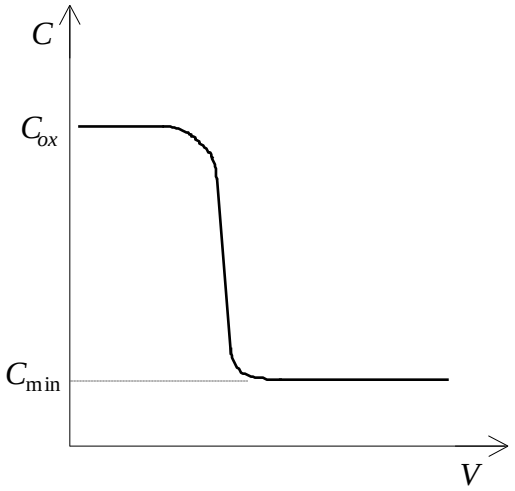

Fig. 42: Idealized high frequency CV plot (*p*-type substrate; accumulation at left)

Therefore, for high frequency conditions, the capacitance per unit area measured in inversion is the series combination of oxide capacitance per unit area and capacitance per

unit area of the depletion region. Furthermore, since, the depletion width reaches a maximum value, the combined capacitance per unit area saturates at C_{min}.

Interpretation of Ideal Capacitance-Voltage Measurements

In principle, measurements of capacitance versus voltage can be made either by sweeping the applied voltage from accumulation to inversion (– to + voltage for *p*-type; + to – for *n*-type) or inversion to accumulation (+ to – voltage for *p*-type; – to + for *n*-type). For quasistatic measurements the direction of the sweep makes essentially no difference in the form of the CV plot, again, since the MOS capacitor remains nearly at equilibrium. However, for high frequency measurements, the form of the CV plot can differ in the inversion region depending on the direction and rate of the voltage sweep. As asserted previously, this behavior is due to the kinetics of minority carrier generation. Clearly, if the time constant for generation-recombination processes is long (indicative of a high quality substrate), then the inversion layer may not fully form during a fast voltage sweep from accumulation to inversion. Thus, the depletion region may grow larger than one would otherwise expect for equilibrium conditions. This phenomenon is called *deep depletion* and is illustrated below:

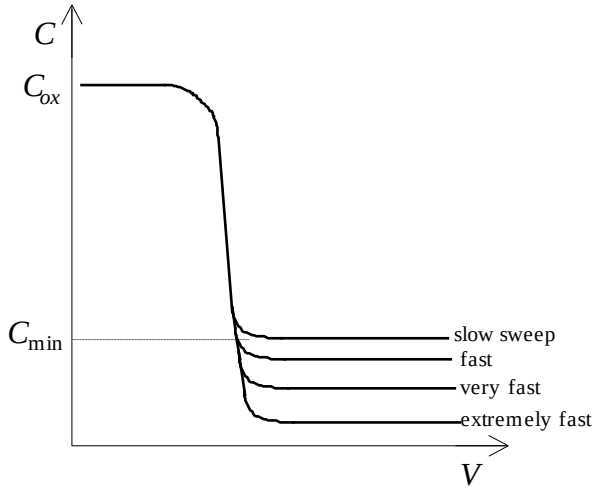

Fig. 43: Deep depletion for sweep from accumulation to inversion (*p*-type substrate; accumulation at left)

Although deep depletion is generally a nuisance in conventional CV analysis and, as such, to be avoided, it is possible to take advantage of this effect to determine the time constant of carrier generation-recombination processes, *i.e.*, minority carrier lifetime, for the substrate. As asserted previously, generation-recombination is slow for a high quality substrate; hence, minority carrier lifetime is long. However, minority carrier lifetime is significantly shortened (on the order of microseconds) in defected or contaminated semiconductor due enhancement generation-recombination processes. Moreover, if minority carrier lifetime is short, not only is deep depletion absent, but, even at high frequency one may observe the onset of equilibrium behavior, *i.e.*, MOS capacitance increases in inversion. Accordingly, minority carrier lifetime can be estimated by determining the dependence of inversion capacitance on sweep rate. The derivative of

this function extrapolated back to equilibrium conditions is proportional to minority carrier lifetime.

In practice, to avoid deep depletion, high frequency CV measurements are nearly always made by sweeping the applied bias voltage from inversion to accumulation. Furthermore, prior to application of the voltage sweep, the substrate is fully inverted by appropriate biasing and illumination of the surface. (Illumination enhances the formation of an inversion layer by providing photo-generated minority carriers.) Of course, the voltage sweep itself should be made without illumination. One finds that the form of associated CV plots essentially depends on oxide thickness and the substrate doping (among other factors). Obviously, ideal MOS capacitance per unit area, *C*, is constructed by the usual series combination:

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_d} = \frac{x_o}{\varepsilon_{ox}} + \frac{x_d}{\varepsilon_s}$$

(In accumulation, x_d vanishes, in inversion, x_d is equal to $x_d^{\max}$, and in depletion x_d varies smoothly between these two values.) Clearly, if substrate doping is constant, an increase in oxide thickness reduces the capacitance of the oxide layer; thus, the total MOS capacitance is also reduced. Furthermore, the position of the depletion region as a function of bias moves to higher values of voltage magnitude since it is really the electric field magnitude at the interface which determines surface conditions (*i.e.*, a higher voltage magnitude must be applied across a thicker oxide to obtain the same electric field magnitude at the Si/SiO_2 interface.) Conversely, if oxide thickness is constant, a change in substrate doping causes a corresponding change in depletion layer capacitance. This is easy to understand because increasing (or decreasing) substrate doping causes the maximum width of the depletion layer to be reduced (or increased). Therefore, for a specified oxide thickness, C_{ox} remains constant and $C_{\min}$ changes as a function of substrate doping; however, the position of the onset of depletion as a function of voltage remains unaffected.

The Effect of Fixed Charges

Ideally, there is essentially no uncovered charge within a high quality thermal oxide layer. However, in reality, it is possible for charges to become more or less permanently trapped within the oxide layer or at the Si/SiO_2 interface. Furthermore, in many cases these charges behave as if they are fixed. Therefore, in analogy to ionized impurity atoms in the substrate, such fixed oxide charges do not participate in nor are changed by charging the MOS capacitor. However, the existence of extraneous fixed charges does cause an overall shift in the position of the depletion region with respect to applied bias voltage. This is easily understood in elementary terms, since if one solves Poisson's equation, one finds that a layer of fixed charge inside the oxide layer just results in a constant potential offset. This is most conveniently analyzed by considering the capacitance and voltage for which the semiconductor is in a flat band condition.

Accordingly, one begins by considering surface differential capacitance per unit area in the semiconductor substrate subject to the assumption that the MOS capacitor is

essentially at equilibrium. Obviously, this limits consideration to low frequency or quasistatic conditions. Clearly, for an arbitrary surface potential, *i.e.*, arbitrary bias voltage, it follows from the fundamental definition of depletion layer capacitance that:

$$|\delta Q_s| = C_d |\delta \varphi_s|$$

Here, δQ_s and $\delta\varphi_s$ denote differential changes in surface charge density and potential. (Absolute values appear so that this analysis can be applied to either *n* or *p*-type substrates.) Clearly, the partial derivative of surface charge density with respect to dimensionless surface potential is directly obtained from the explicit expression for Q_s constructed previously:

$$\left|\frac{\partial Q_s}{\partial \phi_s}\right| = \left(\frac{\varepsilon_s kT}{q\lambda_i}\right) \frac{|\sinh\phi_s - \sinh\phi_\infty|}{\sqrt{2((\phi_\infty - \phi_s)\sinh\phi_\infty - \cosh\phi_\infty + \cosh\phi_s)}}$$

Clearly, one determines capacitance per unit area directly by multiplying the preceding expression by q/kT, hence:

$$C_d = \left(\frac{\varepsilon_s}{\lambda_i}\right) \frac{|\sinh\phi_s - \sinh\phi_\infty|}{\sqrt{2((\phi_\infty - \phi_s)\sinh\phi_\infty - \cosh\phi_\infty + \cosh\phi_s)}}$$

It is convenient to define a dimensionless band bending potential, ϑ_s, as $\phi_s - \phi_\infty$, hence:

$$C_d = \left(\frac{\varepsilon_s}{\lambda_i}\right) \frac{|\sinh(\vartheta_s + \phi_\infty) - \sinh\phi_\infty|}{\sqrt{2(\cosh(\vartheta_s + \phi_\infty) - \cosh\phi_\infty - \vartheta_s \sinh\phi_\infty)}}$$

Of course, the flat band condition occurs when ϑ_s exactly vanishes; hence, one must consider the limit:

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{|\sinh(\vartheta_s + \phi_\infty) - \sinh\phi_\infty|}{\sqrt{2(\cosh(\vartheta_s + \phi_\infty) - \cosh\phi_\infty - \vartheta_s \sinh\phi_\infty)}}$$

Here, C_d^{FB} is defined as the value of C_d at flat band conditions. To determine the limit, one formally substitutes exponentials for hyperbolic functions

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{\left|e^{\vartheta_s + \phi_\infty} - e^{-\vartheta_s - \phi_\infty} - 2\sinh\phi_\infty\right|}{2\sqrt{e^{\vartheta_s + \phi_\infty} + e^{-\vartheta_s - \phi_\infty} - 2\cosh\phi_\infty - 2\vartheta_s \sinh\phi_\infty}}$$

Next, one makes use of the Taylor series of the exponential function such that:

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{\left|(1+\vartheta_s+\cdots)e^{\phi_\infty} - (1-\vartheta_s+\cdots)e^{-\phi_\infty} - 2\sinh\phi_\infty\right|}{2\sqrt{(1+\vartheta_s+\tfrac{1}{2}\vartheta_s^2+\cdots)e^{\phi_\infty} + (1-\vartheta_s+\tfrac{1}{2}\vartheta_s^2+\cdots)e^{-\phi_\infty} - 2\cosh\phi_\infty - 2\vartheta_s\sinh\phi_\infty}}$$

Obviously, high order terms are negligible in the series expansions; hence, only low order terms have been retained, *viz.*, linear in the numerator and quadratic in the denominator. Thus, it follows immediately from the elementary relationship of exponential and hyperbolic functions that:

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{|2\vartheta_s \cosh\phi_\infty|}{2\sqrt{\vartheta_s^2 \cosh\phi_\infty}} = \left(\frac{\varepsilon_s}{\lambda_i}\right)\sqrt{\cosh\phi_\infty}$$

Likewise, it follows immediately from the definition of Fermi potentials that:

$$\cosh\phi_\infty = \frac{p(\infty)+n(\infty)}{2n_i}$$

Furthermore, in an extrinsically doped semiconductor, majority carriers predominate, hence, one can approximate the preceding expression explicitly in terms of net doping density:

$$\cosh\phi_\infty = \frac{|N_A - N_D|}{2n_i}$$

Thus, the capacitance of the depletion layer per unit area subject to flat band conditions has the form:

$$C_d^{FB} = \frac{\varepsilon_s}{\lambda_i}\sqrt{\frac{|N_A - N_D|}{2n_i}} = \frac{\varepsilon_s}{\lambda_D}$$

Accordingly, *extrinsic Debye length* is defined as follows:

$$\lambda_D = \sqrt{\frac{\varepsilon_s kT}{q^2|N_A - N_D|}}$$

Physically, Debye length is a characteristic distance that some external electric field can penetrate a neutral semiconductor surface without substantially perturbing the semiconductor away from neutrality. (This external field can arise either from a contact potential or an applied bias.) Thus, Debye length can be interpreted as an "effective shielding distance". In many physical systems multiple Debye lengths can be identified. (Indeed, for a semiconductor, both intrinsic and extrinsic Debye lengths are commonly defined.) Moreover, it is evident that the shortest Debye length can be expected to dominate (provided that it is substantially shorter than the nearest alternative). Indeed,

for a typical extrinsic semiconductor at room temperature, $\lambda_D << \lambda_i$. Thus, it is reasonable that C_d^{FB} is simply ε_s/λ_D. Obviously, flat band capacitance per unit area, C_{FB}, is readily constructed, thus:

$$\frac{1}{C_{FB}} = \frac{1}{C_{ox}} + \frac{1}{C_d^{FB}} = \frac{x_o}{\varepsilon_{ox}} + \frac{\lambda_D}{\varepsilon_s}$$

As expected, C_{FB} appears as a series combination of oxide and depletion layer capacitances.

If substrate doping is known, C_{FB} is easily determined from measured values of C_{ox} and $C_{\min}$ obtained from a high frequency CV plot. First of all, one observes that maximum depletion width and extrinsic Debye length are related directly as follows:

$$x_d^{\max} = 2\lambda_D \sqrt{\ln\left(\frac{|N_A - N_D|}{n_i}\right)}$$

Thus, C_{FB} takes the form:

$$\frac{1}{C_{FB}} = \frac{1}{C_{ox}} + \frac{1}{2C_s\sqrt{\ln(|N_A - N_D|/n_i)}}$$

Naturally, C_s is directly related to $C_{\min}$ as follows:

$$\frac{1}{C_s} = \frac{1}{C_{\min}} - \frac{1}{C_{ox}} = \frac{1}{C_{ox}}\left(\frac{C_{ox}}{C_{\min}} - 1\right)$$

Therefore, it immediately follows that:

$$\frac{1}{C_{FB}} = \frac{1}{C_{ox}}\left(1 + \frac{(C_{ox}/C_{\min}) - 1}{2\sqrt{\ln(|N_A - N_D|/n_i)}}\right)$$

Of course, this expression is easily recast in terms of absolute flat band capacitance, C_{FB}, and measured values of absolute capacitances, C_{max} and C_{min}, again, obtained from a high frequency CV plot:

$$C_{FB} = \frac{C_{\max}}{1 + \frac{(C_{\max}/C_{\min}) - 1}{2\sqrt{\ln(|N_A - N_D|/n_i)}}}$$

This formula frequently appears in practical guides to CV measurements. Clearly, C_{FB} is the series combination of capacitances, $C_{ox}A$ and $C_d^{FB}A$. Furthermore, once C_{FB} has been

determined, flat band voltage, V_{FB}, is easily specified from experimental data (*i.e.*, one just "reads off" the voltage that corresponds to C_{FB} from the CV plot).

Ideally, the flat band voltage should correspond just to the effective work function difference between the metal contact and the doped silicon substrate. However, if uncovered charges are present within the oxide layer, then the flat band voltage corresponds to the expression:

$$V_{FB} = \phi_M - \phi_{Si} - \frac{1}{\varepsilon_{ox} A} \int_0^{x_o} dx' x' \rho_{ox}(x')$$

Here, $\rho_{ox}(x)$ is charge density in the oxide and is regarded as a function of depth from the oxide surface. (This expression is easily derived from Poisson's equation.) Clearly, the nearer a charge is the Si/SiO_2 interface, the larger is its contribution to flat band voltage. In many cases, it is reasonable to assume that all of the oxide charge is located at or very near the Si/SiO_2 interface. In this case, one defines fixed charge per unit area, Q_f; hence, the above expression takes the form:

$$V_{FB} = \phi_M - \phi_{Si} - \frac{x_o Q_f}{\varepsilon_s} = \phi_M - \phi_{Si} - \frac{Q_f}{C_{ox}}$$

Almost invariably, Q_f is found to be positive, in which case V_{FB} is more negative than the work function difference. Hence, if positive fixed charges are present near the Si/SiO_2 interface, then the CV plot is translated to more negative values of bias voltage, but functional form remains undistorted. This translation from the ideal flat band voltage corresponding to the simple work function difference, to flat band voltage experimentally observed in a CV plot is called *flat band shift*, ΔV_{FB}, and is illustrated below:

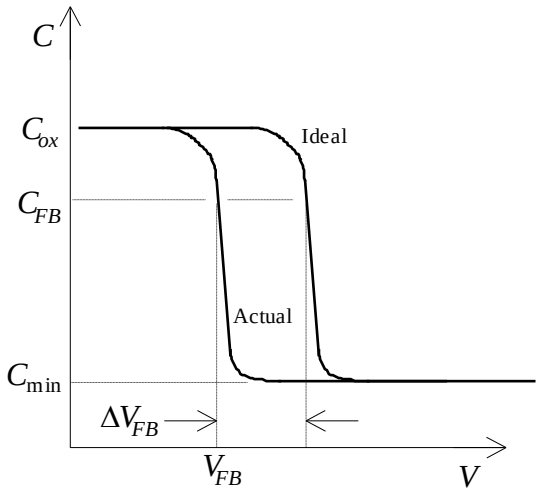

Fig. 44: Flat band shift due to oxide fixed charge (*p*-type substrate; accumulation at left)

Clearly, the actual CV plot is shifted by ΔV_{FB} toward more negative bias voltage in comparison to the ideal CV plot. The flat band shift, ΔV_{FB}, has a magnitude of Q_f / C_{ox} and, therefore, is a direct measure of fixed charge density.

Physically, fixed charges arise from a variety of sources. They may be "grown in" due to the oxidation process itself or they may be the result of contamination or radiation damage. Of particular interest are fixed charges that arise as a result of mobile ion contamination (K^+, Na^+, Li^+, *etc.*). Mobile ion contamination may be differentiated from other types of fixed charge by use of a *bias stress test*. The technique is as follows: First, an initial high frequency CV measurement is made. Second, a large positive bias is applied to the gate and simultaneously the substrate is heated to ~200°C. This treatment should serve to sweep any mobile ions dissolved in the oxide to the Si/SiO_2 interface. Third, the bias is removed, the substrate cooled, and a second CV measurement is made. If the second CV plot is translated to more negative values in comparison to the initial CV plot, but there is no significant distortion in the shape, then the presence of mobile ion contamination within the oxide layer is indicated. This result can be confirmed by applying a large negative bias to the gate and, again, heating the substrate. A CV plot obtained after this treatment should show some evidence of "recovery", *i.e.*, translation of the CV plot back to more positive values.

The Effect of Interface Traps and Fast Surface States

Other important phenomena readily observable in CV measurements are the existence of interface traps and fast surface states at the Si/SiO_2 interface. Theoretically, such interface quantum states must arise naturally due to the broken symmetry caused by termination of the crystal lattice at a surface. Furthermore, these states may have characteristic energies that lie within the band gap of the bulk crystal, thus, reducing minority carrier lifetime. Physically, interface traps and fast surface states can be regarded as arising from unsatisfied, *i.e.*, dangling, bonds which appear upon transition from single crystal silicon to amorphous silicon dioxide. Generally, it is found that the density and energy distribution of interface traps and fast surface states is very dependent on processes and/or materials. Semantically, the difference between interface traps and fast surface states is somewhat vague; however a useful distinction can be made by consideration of time constants. Within this context, a fast surface state has a charging time constant sufficiently short so that its response is observable in a high frequency CV measurement. This causes distortion and "stretch-out" of the corresponding CV plot. In contrast, an interface trap has a longer charging time constant so that distortion of high frequency CV measurements is minimal. Of course, both fast surface states and interface traps are observable in quasistatic CV measurements. Obviously, if significant distortion is visible in a measured high frequency CV plot, then the quality of the Si/SiO_2 interface is very poor and no further analysis is required. Such distortion is illustrated in the following figure

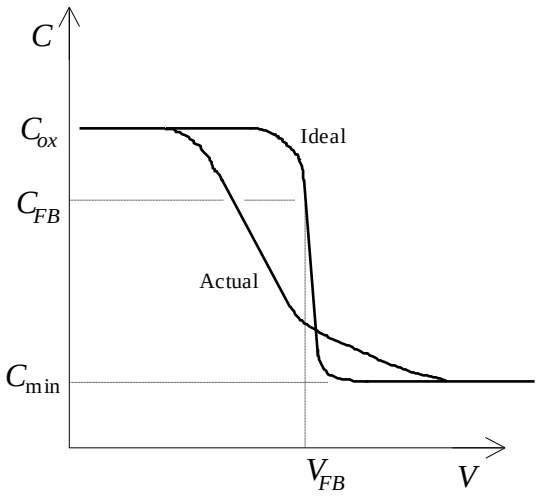

Fig. 45: Distortion due to fast surface states (*p*-type substrate; accumulation at left)

As a practical matter, it has long been known that fast surface states can be "passivated" by means of low temperature heat treatment (400-450°C) in a hydrogen containing ambient, *e.g.*, in "forming gas". (Presumably, the hydrogen diffuses to the Si/SiO_2 interface and satisfies dangling bonds.)

However, if high frequency CV measurements exhibit minimal distortion, then measurement of interface trap capacitance per unit area, C_{it}, is particularly useful as a quantitative measure of the quality of a Si/SiO_2 interface. In practice, C_{it} can be determined by a direct comparison of quasistatic and high frequency CV data taken for the same capacitor. Obviously, the essence of the method relies on the charging kinetics of interface traps. Again, in a quasistatic measurement, interface traps result in a contribution to capacitor charging current, but since time constants of interface traps are relatively long, the effect of interface traps is absent (or at least greatly reduced) in a high frequency measurement. Therefore, the difference between quasistatic and high frequency CV measurements under conditions of depletion allows direct determination of C_{it}:

$$C_{it} = C_d^{LF} - C_d^{HF} = \left(\frac{1}{C_{LF}} - \frac{1}{C_{ox}}\right)^{-1} - \left(\frac{1}{C_{HF}} - \frac{1}{C_{ox}}\right)^{-1}$$

Here, C_d^{LF} and C_d^{HF} are depletion layer capacitances per unit area obtained respectively by observation of quasistatic and high frequency MOS capacitances per unit area, C_{LF} and C_{HF}. Clearly, in the preceding formulation, observed depletion layer capacitance is regarded as the series combination of interface trap capacitance and "true" depletion layer capacitance. Of course, absolute interface trap capacitance, C_{it}, is just $C_{it}A$ and, thus, corresponds to the formula:

$$C_{it} = \left(\frac{1}{C_{LF}} - \frac{1}{C_{max}}\right)^{-1} - \left(\frac{1}{C_{HF}} - \frac{1}{C_{max}}\right)^{-1}$$

Obviously, C_{LF} and C_{HF} are observed quasistatic (*i.e.*, low frequency) and high frequency MOS capacitances and, naturally, C_{max} is just $C_{ox}A$. Alternatively, one can define ΔC simply as the difference between high frequency and low frequency capacitances, *i.e.*, $C_{LF}-C_{HF}$:

$$C_{it} = C_{max}\left(\left(\frac{C_{max}}{C_{HF}+\Delta C}-1\right)^{-1}-\left(\frac{C_{max}}{C_{HF}}-1\right)^{-1}\right) = \Delta C\left(1-\frac{C_{HF}+\Delta C}{C_{max}}\right)^{-1}\left(1-\frac{C_{HF}}{C_{max}}\right)^{-1}$$

Of course, it is desirable for C_{it} to be small.

To express these measurements in a more fundamental form, it is possible to relate C_{it} directly to total interface trap density, D_{it}. To begin, one observes that interface traps may exhibit either acceptor-like or donor-like behavior. Physically, this means that acceptor-like traps behave in analogy to shallow level impurity states associated with acceptor dopant atoms and, hence are negatively charged when occupied by electrons. Conversely, donor-like trap states behave similarly to shallow level impurity states arising from donor dopant atoms and are positively charged when occupied by holes, *i.e.*, unoccupied by electrons. Of course, in contrast to shallow level impurity states, which are distributed throughout the bulk of the semiconductor crystal, interface trap states are localized at the Si/SiO_2 interface. Physically, at equilibrium interface trap charge per unit area due to acceptor-like traps, Q_{it}^{a}, corresponds to the integral expression:

$$Q_{it}^{a} = -q\int_{E_V}^{E_C} dEf(E-q\varphi_s)D_{it}^{a}$$

Here, D_{it}^{a} is the density of acceptor-like interface trap states and $f(E)$ is the Fermi-Dirac distribution function for electrons. In this analysis the rigorous Fermi-Dirac distribution function must be used since interface trap states may have energies close to the Fermi level, *i.e.*, inside the band gap. Of course, an explicit factor of $-q$ must also appear since electrons carry a single negative fundamental unit of charge. Naturally, a complementary expression for interface trap charge per unit area due to donor-like traps, Q_{it}^{d}, can be also be constructed as a definite integral as follows:

$$Q_{it}^{d} = q\int_{E_V}^{E_C} dE(1-f(E-q\varphi_s))D_{it}^{d}$$

Of course, D_{it}^{d} is the density of donor-like interface trap states and in this case, the factor q appears since holes are positively charged. Furthermore, $1-f(E)$ appears within the integrand instead of $f(E)$ since the charged state of a donor-like trap corresponds to an unoccupied electronic state. In both of the preceding expressions, the integral is taken over energies within the band gap.

Of course, during a CV measurement, bias voltage is slowly varied, which causes a corresponding variation in the surface potential, $\delta\varphi_s$. Therefore, one can write:

$$\delta Q_{it} = \delta Q_{it}^d + \delta Q_{it}^a = q\delta\varphi_s \frac{\partial}{\partial\varphi_s} \int_{E_V}^{E_C} dE((1 - f(E - q\varphi_s))D_{it}^d - f(E - q\varphi_s)D_{it}^a)$$

Here, δQ_{it}^a and δQ_{it}^d are variations in interface trap charge associated with acceptor-like and donor-like states, respectively, and δQ_{it} is the variation in total interface trap charge. By definition, $q\varphi_s$ is the difference between the actual Fermi level and the intrinsic Fermi level at the Si/SiO_2 interface. Therefore, it is clear that if φ_s becomes more negative, acceptor-like states discharge (become neutral) and donor-like states become charged. Conversely, if φ_s becomes more positive, acceptor-like states become negatively charged and donor-like states discharge. Since, capacitance is always defined to be positive, it follows that C_{it} is merely $-\delta Q_{it}/\delta\varphi_s$, hence:

$$C_{it} = q \int_{E_V}^{E_C} dE \frac{\partial f(E - q\varphi_s)}{\partial\varphi_s} D_{it}$$

Here, D_{it} is the total interface trap density and is the sum of D_{it}^a and D_{it}^d. As a matter of mathematics, the partial derivative with respect to φ_s can be formally replaced with a partial derivative with respect to E:

$$C_{it} = -q^2 \int_{E_V}^{E_C} dE \frac{\partial f(E - q\varphi_s)}{\partial E} D_{it}$$

From the well-known form of the Fermi-Dirac distribution function, an explicit expression for the partial derivative immediately follows:

$$\frac{\partial}{\partial E} f(E - q\varphi_s) = \frac{-e^{(E - q\varphi_s - E_F)/kT}}{kT(1 + e^{(E - q\varphi_s - E_F)/kT})^2} = -\frac{1}{kT} f(E - q\varphi_s)[1 - f(E - q\varphi_s)]$$

This function is sharply peaked about a value of E equal to $q\varphi_s$ with a peak width of order kT. Therefore, a good approximation to the exact integral is obtained if one assumes that D_{it} is constant over an energy interval of order kT, *i.e.*, one assumes that D_{it} is a smooth, slowly varying function of energy, thus:

$$C_{it} \cong -q^2 D_{it} \int_{E_V}^{E_C} dE \frac{\partial f(E - q\varphi_s)}{\partial E} = q^2 D_{it}(f(E_V - q\varphi_s) - f(E_C - q\varphi_s))$$

Provided that φ_s is has a value that does not locate the Fermi level near the band edges, it is obvious that the Fermi-Dirac distribution function difference is very close to unity, hence:

$$C_{it} \cong q^2 D_{it}$$

It is clear from this expression that the physical interpretation of D_{it} is the number of trap quantum states per unit energy per unit area defined on an energy domain characteristic of the band gap.

A typical plot of D_{it} with respect to surface state energy is shown in the following figure:

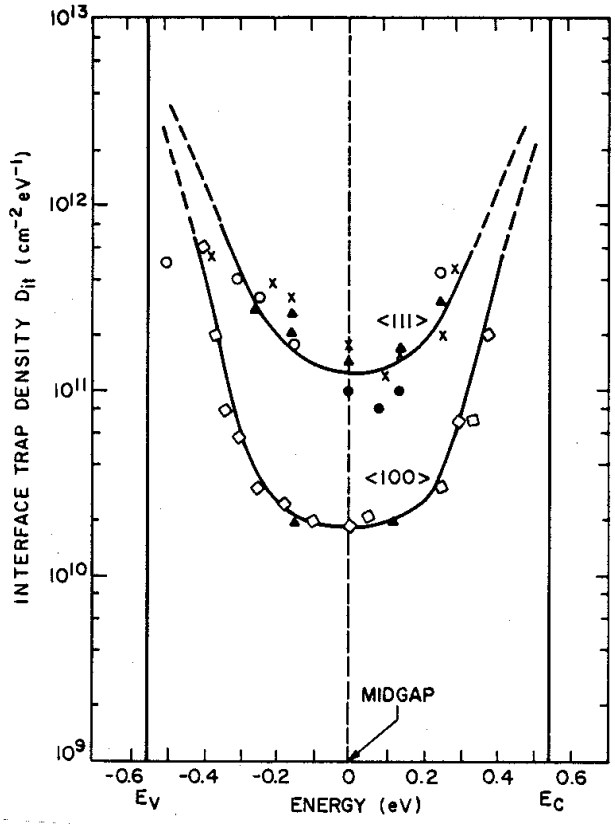

Fig. 46: Interface state density as a function of electronic energy for [111] and [100] silicon surfaces

Clearly, C_{it} and, hence, D_{it} are functions of bias voltage applied during quasistatic and high frequency CV measurements. If one recalls the position of the Fermi level at the semiconductor surface for the various conditions of bias, a physical interpretation of this behavior is readily formulated. Accordingly, irrespective of substrate doping, if the applied bias of an MOS capacitor is swept from negative to positive values, then the surface Fermi level effectively moves upward through the band gap due to changes in band bending. By definition, electronic energy states below the Fermi level tend to be occupied, while those above the Fermi level tend to be empty. Therefore, as the Fermi

level moves upward through the band gap, empty electronic states, *i.e.*, interface trap states and fast surface states, of corresponding energy become occupied. Since electrons carry one fundamental unit of charge, these transitions make a directly observable contribution to C_{it}. The bias voltage at which an interface trap becomes charged is directly related to its energy and, hence, its position relative to the band gap. Thus, observation of C_{it} allows D_{it} to be determined as a function of electronic energy measured relative to the band gap as is shown in the preceding figure. Within this context, a common "rule of thumb" is that D_{it} should no more than $1(10^{11})$ cm^{-2} eV^{-1}. Clearly, this is easily realized on a [100] silicon surface, but not on a [111] surface. (Since, a stable switching threshold is essential to reliable operation of modern CMOS transistors, it is critical to obtain a low density of interface traps; hence, this provides a major motivation for the usual preference of [100] silicon substrates for device fabrication.)