Current-Voltage Measurement

Of course, if a dielectric thin film, e.g., thermal silicon dioxide, is to serve as a high quality insulator, then as asserted previously, it is desirable that very little current, i.e., leakage current, should flow if the insulator layer is subjected to normal bias conditions. Therefore, in addition to desirable capacitance-voltage (CV) characteristics, a thermal oxide layer must also have desirable current-voltage (IV) characteristics. Obviously, these can also be measured using the same MOS structures as fabricated for CV measurements. However, physical interpretation is substantially simpler than in the case of CV analysis since the semiconductor substrate, provided that it is sufficiently conductive, is unimportant in IV characterization of a thermal oxide layer.

In principle, IV measurement is quite simple and simply requires biasing the oxide layer at a designated voltage and then measuring resultant current flow. In practice, one observes the measured current as a function of bias voltage. However, this is exactly what was done in a quasi-static CV measurement. So, what is the difference between IV and quasi-static CV measurements? To be specific, there are two major differences. First of all, for a quasi-static CV measurement, transient displacement current, i.e., capacitor charging current, is measured. In contrast, for an IV measurement, one is interested only in steady-state (DC) current, i.e., true conduction current, which continues to flow after the transient has decayed, i.e., after the MOS capacitor has become fully charged at the applied bias voltage. (Indeed, as asserted previously, this current must be formally subtracted from displacement current in order to obtain an accurate quasi-static CV plot.) Second, the bias voltage in an IV measurement is generally carried to values for which the oxide layer breaks down, i.e., fails as an insulator. Moreover, once breakdown has occurred, the oxide layer is permanently damaged, thus IV measurements can be made only once on any particular MOS structure, i.e., IV testing is essentially destructive. This requires that a substantial number of test structures must be measured in order to generate meaningful statistics. Clearly, breakdown of the oxide at the outset caused by application of a large initial bias voltage must be avoided. Therefore, IV measurements should be made by sweeping voltage bias slowly from zero toward either positive or negative values, but not as in a CV measurement by sweeping from positive to negative voltages through zero bias (or the reverse).

Conduction Mechanisms

For relatively thick oxide layers, the IV response of “good” oxide is quite simple. At voltage biases well below breakdown, very little current flows (on the order of a few pA/cm²). However, once breakdown occurs, current rises very rapidly. On a semi-logarithmic plot (i.e., current plotted on a logarithmic scale versus voltage or electric field plotted on a linear scale) an ideal thick oxide IV response appears as a flat or slowly rising curve below breakdown at which point the plot becomes essentially vertical. Obviously, if a large amount of current flows at voltage biases significantly lower than the expected breakdown voltage, then the oxide quality is poor. Of course, breakdown voltage must be directly dependent on oxide film thickness; however breakdown field strength is essentially independent of thickness. In general, breakdown fields for high quality oxides are of magnitude 10-12 MV/cm.
For thin oxides, IV response is complicated by the phenomenon of quantum mechanical tunneling (which is a fundamental physical phenomenon that is a direct consequence of the Heisenberg Uncertainty Principle). To be specific, it is impossible to confine particles (such as electrons) completely by a finite potential barrier (such as provided by a layer of thermal oxide in an MOS structure). Therefore, some current can always be expected to “leak”, even through a materially “perfect” insulator completely free of defects. Within this context, in a typical plot of IV response, tunneling current appears as a rising characteristic in the region just below breakdown. However, particle confinement substantially depends on barrier thickness, and thus, is significant only for very thin oxide layers. This is apparent in the following figure, which illustrates typical IV responses for “thick”, “thin”, and “very thin” oxide layers:

![Graph showing log J vs. 10 MV/cm for thick, thin, and very thin oxide layers.]

Here, for simplicity break down field strength is taken to be the same for each oxide layer; however in actual practice it is likely to be somewhat more variable. In general, tunneling current is essentially independent of oxide quality.

Physically, it is found that at low bias voltages, several electrical conduction mechanisms can exist within thermal oxide. Naturally, the presence of foreign impurities can greatly enhance conduction due to the introduction of physical defects (pinholes, etc.). For impurity free oxides, however, Frenkel-Poole emission and Fowler-Nordheim tunneling are the most common low-field conduction mechanisms. As asserted previously, tunneling is a normal phenomenon that cannot be prevented, and which allows charge carriers, e.g., electrons, to pass through a potential barrier even though available energy for such a process is insufficient. (This accounts for the terminology; since such a particle does not “pass over” the barrier, but “tunnels through” the barrier.) Accordingly, it follows directly from elementary quantum mechanics that the current-field characteristic for Fowler-Nordheim tunneling has an athermal exponential characteristic of the form:

\[
J = A_{FN} E^2 \exp \left( -\frac{E_o}{E} \right)
\]
Here, $A_{FN}$ and $E_o$ are characteristic constants, $J$ is leakage current density, and $E$ is applied field strength. (A more fundamental expression for $E_o$ can be written in terms of an oxide barrier height, electron effective mass, the fundamental unit of charge, and Planck’s constant.) Of course, Fowler-Nordheim tunneling is important in high quality oxides only if they are very thin. However, if there are a large number of “trap states” for electrons distributed within the oxide layer, then tunneling can occur “trap-to-trap”. This mechanism causes a dramatic increase in leakage current in comparison to oxides which have a low trap state density, \textit{i.e.}, that are essentially trap free. Furthermore, trap states are generally associated with the fixed oxide charges that are observable in CV measurements. Of course, a large amount of oxide fixed charge and an associated high density of trap states is indicative of a poor quality oxide. Therefore, a large amount of oxide leakage current observed at a low bias voltage can be expected to be correlated with a high trap state density.

Frenkel-Poole emission is also mediated by electronic trap states and occurs if the electric field within the oxide layer becomes large enough so that electrons trapped within the oxide layer are directly injected into the conduction band of the semiconductor. Consequently, it is found that the current-field characteristic for Frenkel-Poole emission has the form:

$$J = A_{FP}E \exp\left[-q\left(\phi_B - \sqrt{\frac{qE}{\pi \varepsilon_{ox}}}\right)/kT\right]$$

Here, $A_{FP}$ is a “pre-exponential” constant and $\phi_B$ is a barrier height characteristic of the oxide trap states. Clearly, in contrast to tunneling, Frenkel-Poole emission is a thermally activated process. Another less important conduction mechanism similar to Frenkel-Poole emission is \textit{Schottky emission}, which also has a thermally activated current-field characteristic:

$$J = A^*T^2 \exp\left[-q\left(\phi_B - \sqrt{\frac{qE}{4\pi \varepsilon_{ox}}}\right)/kT\right]$$

Here, $A^*$ is a coefficient known as “effective Richardson constant” and $\phi_B$ is, again, trap barrier height. Both Frenkel-Poole and Schottky emission processes should be negligible in high quality thermal oxide, since both mechanisms require a reasonably large density of trap states within the oxide. Typically, a large density of trap states is the result of contamination, damage, and/or generally poor processing.

Two other possible oxide conduction mechanisms are \textit{ohmic} and \textit{ionic} conduction. Both of these are thermally activated:

$$J = A_\varepsilon E \exp[-q\Delta E_{ae}/kT]$$

$$J = \frac{AE}{T} \exp[-q\Delta E_{ai}/kT]$$
Clearly, these are just Arrhenius forms defined such that $\Delta E_{ae}$ and $\Delta E_{ai}$ are, respectively, activation energies for ohmic and ionic conduction processes and $A_e$ and $A_i$ are corresponding pre-exponential factors. Neither one of these conduction mechanisms should ever be observed in high quality thermal oxide. By definition, ionic conduction can only occur if the oxide is greatly contaminated with some type of mobile ionic species, e.g., sodium. Ohmic conduction can occur only if the chemical composition of the oxide layer is disturbed.

For completeness, it is worthwhile to consider conduction mechanisms associated with normal oxide break down. If the electric field becomes very high, the current density through the oxide may become mobility limited.

$$J = \frac{9\varepsilon_0\mu_e^o \nu^2}{8\varepsilon_o}$$

Clearly, in this expression the electric field does not itself appear, but rather $J$ depends on bias voltage directly. Furthermore, $\mu_e^o$ is identified as electronic mobility within the oxide layer and is analogous to electronic mobility as defined for the semiconductor substrate. Physically, just as in the semiconductor substrate itself, mobility is determined by electron scattering from atomic species, i.e., silicon and oxygen atoms. Of course, since silicon dioxide is an insulator, one expects that $\mu_e^o$ should be much smaller than the corresponding electronic mobility of silicon. If the electric field strength due to the applied bias voltage is relatively small, electron scattering processes are essentially elastic and cause no changes in the oxide network structure. However, at or near break down, the electric field strength becomes large. In this case, free electrons injected into the surface of the oxide layer collide with bound atomic electrons causing them also to become free and then to be accelerated by the applied bias. This process is called electron impact ionization. The newly freed electrons can then collide with additional bound atomic electrons, thus multiplying the current in a “chain reaction” or avalanche. This accounts for the rapidly rising current-field characteristic typical of oxide break down. In these circumstances it is not surprising that the high current density associated with avalanche break down permanently damages the oxide layer.

**Oxide Reliability**

In addition to determination of interface trap density, leakage current, or break down field, an additional critical criterion for oxide quality is reliability, which is quantified as an estimate of expected performance of an oxide layer over some projected usable lifetime. Common methods for determination of oxide reliability are charge-dependent-breakdown (QDB) and time-dependent-breakdown (TDDB) analysis. In QDB analysis, an MOS capacitor is biased using a constant current source. Obviously, as current is “pumped into” the capacitor, bias voltage must rise until, finally, the capacitor breaks down. Total injected charge is determined simply by multiplying the applied current by the time to reach break down. In general, the larger the total injected charge, the more reliable the oxide layer. In contrast, in TDDB analysis, bias voltage is held constant. In
In this case, the current flowing through the capacitor is variable. Again, the MOS structure is subject to electrical stress until break down is observed. In principle, TDDB analysis does not require a particular bias voltage, which may be chosen consistent with device characteristic or simply for convenience. Even so, if the chosen bias is too small, then the length of time to observe break down may become extremely long. Conversely, if the bias voltage is too high, TDDB results may not correspond closely to actual operating conditions. In practice, the bias level for TDDB analysis should be set about twice the maximum bias to which an oxide layer will be subjected during normal operation.

For both QDB and TDDB analysis, results will differ depending on whether electrons are injected into the oxide layer from the substrate or from the gate. Therefore, for purposes of comparison, one must adopt a consistent measurement technique. Also, a reasonably large number of MOS capacitors must be measured to obtain an acceptable degree of statistical confidence. In practice, QDB and TDDB data is interpreted by construction of a cumulative probability plot, which has total injected charge (i.e., QDB) or time (i.e., TDDB) required to observe failure (i.e., oxide break down) as the horizontal axis and fraction failed, i.e., failure probability, as the vertical axis. If, for example, a sample of one hundred MOS capacitors is tested until break down is observed, a cumulative probability plot is constructed by ranking measurements from the smallest observed injected charge or shortest time observed for break down and plotting rank against charge or time. In this case, the rank corresponds directly to probability of failure measured in per cent as illustrated in the following figure:

![Cumulative probability plots showing good reliability, poor reliability, and “infant” mortality](image)

**Fig. 48:** Cumulative probability plots showing good reliability, poor reliability, and “infant” mortality

Obviously, sample size need not restricted to any particular number, e.g., one hundred, but may be chosen arbitrarily provided statistical confidence is sufficient.

A “good” QDB or TDDB result is characterized by a nearly vertical distribution of data points. This indicates that all of the measured structures are very similar in behavior. Naturally, the larger the average total injected charge or the longer the average time required for break down, the more reliable the oxide. In contrast, the data points may be distributed more horizontally over some range of probability. This indicates that the behavior of the measured structures is inconsistent and that the failure probability
distribution is very broad or perhaps even bimodal. Such a result is almost certainly caused by defects or damage in the oxide layer and, in general, represents a “bad” result. Within this context, there is one special case worth consideration, which is characterized by a horizontal distribution of points at low probability above which the points are distributed more vertically and about a reasonably high value of average charge or time. This is characteristic of “infant” mortality. Obviously, it is desirable to eliminate such behavior; however this is not always practical in the fabrication process itself. An alternative (albeit a somewhat costly one) is to perform a “burn-in” in which finished devices are stressed well beyond normal operating conditions. Presumably, this precipitates early or infant failures and the remaining devices should have reliability characteristics of the vertically distributed points.

In general, one observes that absolute QDB and TDDB results will vary with respect to measurement conditions (i.e., injected current density, bias voltage, temperature, etc.). However, QDB and TDDB results observed under different conditions can be directly compared by application of a suitable reliability model. In this case, one obtains a mean time to failure or MTTF extrapolated from actual measurement conditions to some normal operating condition. There are several reliability models available for this purpose and there is still considerable debate regarding which model is more realistic. However, if TDDB measurements are made at several bias voltages, it is straightforward to extrapolate average break down time at a value of electric field consistent with normal operation. Although not always possible, it is desirable that this result, i.e., MTTF, should be quite long (perhaps, even a few hundred years). In practice, the desired extrapolation may be made by plotting average break down time versus electric field or reciprocal field. Physically, use of the reciprocal field is justifiable since the logarithm of Fowler-Nordheim tunneling current density is proportional to $1/E$. However in practice, more realistic estimates of MTTF seem to be obtained from empirical extrapolations using just the electric field itself. Ideally, for QDB measurements, the actual charge required for oxide failure should be independent of the magnitude of forced current. In this case, MTTF can be estimated just from the measured value of QDB and oxide leakage current characteristic of normal operation. However, as a practical matter QDB may be found to depend on forcing current. In this case, an extrapolation of QDB to operating conditions can be made in close analogy to methods for extrapolation of MTTF from TDDB measurements.
Physical Characterization of Thermal Oxide

In addition to electrical characterization of thermal oxide using CV or IV methods, there are other useful physical techniques for characterization of thermal oxide films. These generally rely on optical measurements and are used to measure physical film thickness, refractive index, etc.

Reflectance Spectroscopy and Interferometry

In general, it is well known that a transparent thin film having a thickness commensurate with the wavelength of visible electromagnetic radiation will appear colored when it is illuminated by a broad band white light source, e.g., sunlight or other incandescent source. This is caused by interference between light reflected from the top and bottom interfaces of the thin film. Of course, the intensity of the various reflected spectral components is determined by the relative phase between the two reflections. Naturally, the most intense reflected wavelengths will be those for which reflected components are “in-phase”, i.e., interference is constructive. Consequently, it is evident that not all components of illuminating white light are reflected uniformly and, thus, the reflected light appears colored rather than white. (This same phenomenon is readily observed in everyday life in the colors generated by thin oil or soap films.) The apparent color or more precisely the spectral composition of the reflected light is directly related to the thickness of the thin film. This phenomenon is summarized in the following table:

<table>
<thead>
<tr>
<th>μm</th>
<th>Apparent Color</th>
<th>μm</th>
<th>Apparent Color</th>
<th>μm</th>
<th>Apparent Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>Metallic or white</td>
<td>0.39</td>
<td>Yellow</td>
<td>0.77</td>
<td>Yellow; washed out</td>
</tr>
<tr>
<td>0.05</td>
<td>Tan</td>
<td>0.41</td>
<td>Light orange</td>
<td>0.80</td>
<td>Orange; quite strong</td>
</tr>
<tr>
<td>0.07</td>
<td>Brown</td>
<td>0.42</td>
<td>Carnation pink</td>
<td>0.82</td>
<td>Salmon</td>
</tr>
<tr>
<td>0.10</td>
<td>Dark violet to red-violet</td>
<td>0.44</td>
<td>Violet-red</td>
<td>0.85</td>
<td>Light red-violet; dull</td>
</tr>
<tr>
<td>0.12</td>
<td>Royal blue</td>
<td>0.46</td>
<td>Red-violet</td>
<td>0.86</td>
<td>Violet</td>
</tr>
<tr>
<td>0.15</td>
<td>Light blue to metallic blue</td>
<td>0.47</td>
<td>Violet</td>
<td>0.87</td>
<td>Blue-violet</td>
</tr>
<tr>
<td>0.17</td>
<td>Metallic to very light yellow-green</td>
<td>0.48</td>
<td>Blue-violet</td>
<td>0.89</td>
<td>Blue</td>
</tr>
<tr>
<td>0.20</td>
<td>Light gold or yellow; slight metallic look</td>
<td>0.49</td>
<td>Blue</td>
<td>0.92</td>
<td>Blue-green</td>
</tr>
<tr>
<td>0.22</td>
<td>Gold with slight yellow-orange</td>
<td>0.50</td>
<td>Blue-green</td>
<td>0.95</td>
<td>Yellow-green; dull</td>
</tr>
<tr>
<td>0.25</td>
<td>Orange to melon</td>
<td>0.52</td>
<td>Green; quite strong</td>
<td>0.97</td>
<td>Yellow; somewhat washed out</td>
</tr>
<tr>
<td>0.27</td>
<td>Red-violet</td>
<td>0.54</td>
<td>Yellow-green</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>Blue to violet-blue</td>
<td>0.56</td>
<td>Green-yellow</td>
<td>0.99</td>
<td>Orange</td>
</tr>
<tr>
<td>0.31</td>
<td>Blue</td>
<td>0.57</td>
<td>Yellow; washed out</td>
<td>1.00</td>
<td>Carnation pink</td>
</tr>
<tr>
<td>0.32</td>
<td>Blue to blue-green</td>
<td></td>
<td></td>
<td>1.02</td>
<td>Violet-red</td>
</tr>
<tr>
<td>0.34</td>
<td>Light green</td>
<td></td>
<td></td>
<td>1.05</td>
<td>Red-violet</td>
</tr>
<tr>
<td>0.35</td>
<td>Green to yellow-green</td>
<td></td>
<td></td>
<td>1.06</td>
<td>Violet</td>
</tr>
<tr>
<td>0.36</td>
<td>Yellow-green</td>
<td></td>
<td></td>
<td>1.07</td>
<td>Blue-violet</td>
</tr>
<tr>
<td>0.37</td>
<td>Green-yellow</td>
<td></td>
<td></td>
<td>1.10</td>
<td>Green</td>
</tr>
<tr>
<td>0.39</td>
<td>Yellow</td>
<td></td>
<td></td>
<td>1.11</td>
<td>Yellow-green</td>
</tr>
<tr>
<td>0.40</td>
<td>Yellow</td>
<td></td>
<td></td>
<td>1.12</td>
<td>Green</td>
</tr>
<tr>
<td>0.41</td>
<td>Light orange</td>
<td></td>
<td></td>
<td>1.18</td>
<td>Violet</td>
</tr>
<tr>
<td>0.42</td>
<td>Carnation pink</td>
<td></td>
<td></td>
<td>1.20</td>
<td>Violet-red</td>
</tr>
</tbody>
</table>

Table 3: Apparent colors of thermal oxide of various thickness (in μm; viewed normal to the surface)
Here, oxide thickness is specified in micrometers and apparent colors are specified empirically. Naturally, reflectance spectroscopy allows quantification and relies on analysis of the spectral composition of normally reflected light from an oxide thin film. Indeed, in simplest form no equipment other than a “calibrated human eyeball” is required. As is clear from the preceding table, experienced observers can easily estimate oxide thickness to within a few nanometers. However, in recent years, automated instrumentation has been developed which is both more accurate and convenient than simple visual observation. In addition, these instruments collect reflected light through a microscope objective, which allows thickness measurements to be made at very precise locations on the substrate surface. This is particularly useful for characterization of partially fabricated devices. Indeed, if multiple positions are measured using some predefined pattern that essentially samples the whole wafer, then the data can be conveniently rendered into a “map” of thickness.

If instead of a broad band light source, a monochromatic light source (such as a laser) is used, reflectance spectroscopy becomes reflectance interferometry. Rather than as an “after-the-fact” characterization method, reflectance interferometry is most useful as an “in-situ” measuring technique, which can be incorporated into processing equipment allowing film thickness to be measured during growth or removal (e.g., by etching or polishing).

**Monochromatic and Spectroscopic Ellipsometry**

Ellipsometry is a second optical interferometric technique that is frequently used to characterize transparent thin films, e.g., thermal oxide. In its simplest form, a monochromatic beam of light (typically from a laser diode) is resolved into two independent polarized components. These are reflected from the wafer surface at some fixed angle. A second variable polarizer then analyzes the reflected light. It is well known from classical electromagnetic field theory that light components polarized parallel and perpendicular with respect to the surface exhibit different, independent behavior with respect to reflection. As before for reflectance spectroscopy, interference occurs between light reflected from the top and bottom interfaces of the oxide layer. Moreover, the intensity and phase of each polarized component of the reflected light is characteristically dependent on thickness and refractive index of the thin film. Since, the two polarized components are independent, monochromatic ellipsometry can make simultaneous measurements of both thickness and refractive index of an oxide layer. In addition, ellipsometry is inherently more precise than a reflectance spectroscopy and is particularly useful for characterization of very thin oxide layers. Ellipsometry can be extended either by measuring at various angles of reflection, (multiangle ellipsometry) or by using various wavelengths of light (multiwavelength ellipsometry). In both of these cases, each measurement made at different angles or wavelengths provides two independent values of thickness and refractive index. For a single thin film (such as a thermal oxide layer) these can be used to increase the accuracy of the overall measurement. However, a more common application of these techniques is simultaneous, independent measurement of refractive indices and thicknesses of two or more “stacked” transparent thin film layers.
Extending ellipsometry further, if an intense broad band light source is used, ellipsometric measurements can be made over a continuous range of wavelengths. This is spectroscopic ellipsometry. Clearly, since two independent measurements can be made at any particular wavelength, the amount of information available in a single spectroscopic ellipsometric measurement is quite large. Again, this is useful for characterization of multilayer thin films, however, for characterization thermal oxide spectroscopic ellipsometry can be used to accurately account for the optical properties of the Si/SiO$_2$ interface and/or the semitransparent surface layers of the substrate itself. In practice, this requires extensive numerical fitting to some “model” of the oxide or Si/SiO$_2$ interface. In recent years, several systems have been developed which provide algorithmic support for spectroscopic ellipsometry. This allows very accurate characterization of thin thermal oxide films.

**Prism Coupling**

A third optical technique that can be used for thin film characterization is prism coupling. This method is used less frequently at present than previously since it suffers from the disadvantage that, in contrast to ellipsometry or reflectance spectroscopy which are non-contacting; a small prism must come in contact with the thin film surface. The prism is made of a transparent material chosen so that the interface between the prism and the thin film layer forms a totally reflecting interface. Physically, it turns out that if the prism-thin film couple is rotated with respect to a monochromatic optical source (i.e., a laser) some of the totally reflected light “leaks out” into the thin film due to the phenomenon of evanescent coupling. As might be expected, the degree of evanescent coupling depends on refractive index and thickness of the thin film material and appears as a series of interference fringes that are a function of angular position. In practice, if several interference fringes (more than three or four) can be observed, then both thickness and refractive index of the thin film can be determined. Prism coupling is most applicable to relatively thick transparent thin films. This is useful since both reflectance spectroscopy and ellipsometry typically become less accurate for thick films due to cyclic error, *etc.*
Pre-Oxidation Cleaning

It is critical that prior to any thermal oxidation process, the silicon surface should be scrupulously cleaned. Typically, this is done by treating the wafers in two successive chemical solutions conventionally called “SC-1” and “SC-2” (surface cleans 1 and 2). The precise composition of these solutions is somewhat variable; however, SC-1 is generically formulated as a 10:1 mixture of commercial ammonium hydroxide (NH₄OH) and hydrogen peroxide (H₂O₂) solutions heated to about 80°C. Similarly, SC-2 is also approximately a 10:1 mixture of commercial hydrochloric acid (HCl) and hydrogen peroxide (H₂O₂) solutions, again, heated to about 80°C. Naturally, all starting materials must be “electronic grade”. These mixtures are also sometimes diluted with an equivalent volume of de-ionized water; however, this is not a requirement. It has been shown that SC-1 is effective primarily for removal of organic contamination and SC-2 for removal of metallic species. Therefore, any type of contamination is substantially reduced by sequential treatment in SC-1 and SC-2. Originally, of course, these cleans were carried out in simple static tanks. As particle control has become more critical, recirculation, filtration, and automation have been added. Alternatively, it is common practice to use completely automated spray chemical processors in which wafers are cleaned, rinsed, and dried without any external intervention.

In addition to contamination, it is often desirable to remove any pre-existing “native oxide” from the silicon surface. (Native oxide is a naturally occurring thin oxidized layer one to two nm thick on the surface of silicon that arises from routine exposure to oxygen and water vapor in the atmosphere.) This may be done by a short etch in unbuffered 50:1 hydrofluoric acid solution following cleaning in SC-1 and SC-2. (Buffered oxide etch or BOE should not be used because it contains the salt, ammonium fluoride, which can form particles on the surface.) Removal of native oxide is thought to result in a “hydrogen terminated” silicon surface, which is believed to persist perhaps one to four hours depending on conditions, before the native oxide layer is reformed.

In addition to wet chemical treatments, vapor phase processing using anhydrous hydrogen fluoride is also an option for pre-oxidation cleaning. In recent years, various systems have been developed for this purpose. Some of these have even been integrated with the oxidation process. None of these seem to have proved entirely satisfactory both in terms of cost and performance.
Ultra-Thin Insulators

As will become evident in more detail subsequently, in a very real sense an MOS transistor is the solid-state analog of an old-fashioned vacuum tube. (It is an MOS capacitor built on top of a semiconductor resistor hence, the term “transfer resistor” or transistor.) Obviously, any bias voltage applied to the gate modulates current flow in the channel. Of course, the conduction channel is electrically connected to the circuit wiring on each end (source and drain contacts). Therefore, the gate is insulated and ideally should not supply any current to the channel. Within this context, one observes that state-of-the-art MOS transistors require fabrication of ultra-thin gate insulators which, in addition, must satisfy stringent performance specifications. Indeed, insulating properties must not only be excellent, but since the gate electrode is generally made of heavily doped CVD polysilicon, metal, or metal alloy, the gate insulator must also block any migration of metal atoms or shallow level dopant impurity, in particular boron, from the gate electrode into the conduction channel as well. (If there is any significant dopant contamination of the channel from the polysilicon gate electrode, the transistor threshold voltage will have unacceptable variation.)

Reoxidized Nitrided Oxide

Unfortunately, a pure thermal oxide film has poor characteristics with respect blocking dopant migration from the polysilicon (particularly in the case of boron). This problem may be addressed by direct incorporation nitrogen into a thermal gate oxide film. Within this context, one might ask, why not just use pure silicon nitride? Certainly, nitride is a good insulator and, indeed, very thin nitride layers can be produced by CVD or even direct thermal nitridation of a silicon surface. However, a silicon nitride/silicon interface has a very high density of interface traps. Within the context of MOS transistor performance, this causes a severe degradation of effective carrier mobility in the channel. (In practical terms, this appears as a high channel resistance or low drive current.) As a practical matter, it is found that if the nitrogen concentration exceeds one atomic per cent at the gate oxide/silicon interface there is a significant degradation of device performance. Thus, one would ideally like to have a graded nitrogen concentration in the gate oxide with the concentration relatively high at the polysilicon/gate oxide interface and low at the gate oxide/silicon interface. However, this is difficult to achieve in practice.

The most practical process for formation of a nitrided oxide is the addition of either nitrous (N₂O) or nitric (NO) oxide directly to the oxidizing ambient. Unfortunately, since further oxidation tends to convert nitride to oxide, this also results in the occurrence of maximum nitrogen concentration precisely at the gate oxide/silicon interface. Thus, for a one step oxidation process, the total concentration of nitrogen in the oxide must be kept low. This suggests implementation of a two step process as a possible improvement. In the first step a relatively nitrogen rich oxide is grown. This is followed by “reoxidation” in non-nitrogen containing ambient. Since new oxide is formed only at the gate oxide/silicon interface, the interfacial nitrogen concentration falls rapidly. However, again since oxidation converts nitride to oxide, the nitrogen content of the entire film also falls. Thus, the growth rate of oxide at the interface must be balanced with the
conversion of nitride to oxide in the bulk of the film. Fortunately, this trade-off can be achieved since it is found in practice that only a small nitrogen concentration is effective at blocking “boron penetration”.

In passing, it is worthwhile to mention other alternative approaches for fabrication of nitrided gate oxide. In particular, very shallow ion implantation of nitrogen either into a preformed gate oxide or into the silicon substrate itself upon which a subsequent gate oxide is grown have both been tried. In either case, results do not appear to be as good as that obtained using some form of N₂O or NO oxidation processes.

**Rapid Thermal Oxidation**

Alternatively, rapid thermal oxidation (RTO) can also produce a high quality oxide. Typically, RTO is implemented by use of high intensity quartz-halogen lamps rather than ordinary resistive heating elements as a heat source. Of course, the conventional tube configuration is generally not optimal for RTO, which is much more compatible with “single wafer processing”. Accordingly, individual wafers are typically suspended on rings or pins within a small process chamber to reduce thermal mass and allow rapid change of the temperature. Indeed, it is possible to achieve a very high wafer surface temperature (>1000°C) quite quickly, e.g., in less than a minute. Therefore, in principle, thin gate oxides can be controllably grown in a very short period of time. In practice, RTO processes have suffered from problems of repeatability and control and, thus, have not been used widely for conventional oxidation.

**Post Oxidation Annealing**

As observed previously, for a high quality gate oxide, $D_\alpha$ should be no more than $1(10^{10})$ cm$^{-2}$ eV$^{-1}$. If this is to be achieved in a single oxidation step, this requires oxidation at very high temperature. (Conventionally, this favors dry oxidation over steam since the lower rate allows for better thickness control.) Alternatively, post-oxidation annealing in an inert ambient at high temperature (>1000°C) can reduce an unacceptably high post-oxidation $D_\alpha$ to a desirable value. Presumably, this “repairs” the Si/SiO₂ interface. Of course, it goes almost without mentioning that any high quality oxidation process requires scrupulous pre-cleaning of the substrate surface.

**Limitation of Conventional Oxidation Technology**

The channel length for the current device generation is 20 nm (or less). Such short channel lengths require gate insulator thicknesses equivalent to no more than 1-2 nm of pure silicon dioxide in order to achieve desirable device characteristics. (If the gate insulator thickness is not scaled with the channel length, resulting transistors suffer from severe “short channel effect”.) Fabrication of such thin layers requires careful processing, although, perhaps somewhat surprisingly, process conditions using conventional quartz tube furnaces can be found that result in high yield and very reliable insulating films. In principle, thin oxides can be produced by either diluting the oxidant with inert gas and/or reducing the oxidation temperature. In the latter case, a subsequent high temperature anneal in an inert ambient is needed to reduce interface trap density. In
practice, even high quality thin oxide films generally must be modified or replaced to reduce current flowing between the gate electrode and the channel. Indeed, this is not due to formation of defects in thin oxide films and there is no inherent problem in fabrication of such thin films by thermal oxidation; however, current resulting directly from quantum mechanical tunneling of electrons through the oxide layer can be expected to be impractically large. This presents a very severe limitation on device performance. Therefore, either physical characteristics of thermal oxide must be markedly improved (not likely), a new device architecture must be invented (e.g., the recently introduced “tri-gate” structure), or some other methodology must be found.

Obviously, one solution to the problem of thin gate insulators is partial or complete replacement of thermal silicon dioxide with some other material. Any such material must have a large dielectric constant, i.e., “high-k”, such that physical thickness can be much larger than the physical thickness of an electrically equivalent silicon dioxide layer. These are generally identified with oxides and/or silicates of heavy metals such as zirconium, tantalum, or hafnium. (Other possibilities include a host of perovskite materials.) Of course, as one might imagine such a radical change in device structure is not to be undertaken unless out of absolute necessity since it is precisely the unique compatibility of the silicon-silicon dioxide material system that makes modern solid-state electronics really possible. In particular, for any insulator different from thermal oxide a much poorer silicon/insulator interface, viz., high trap density, is to be expected. Even so, in recent years these kinds of changes have been required for the most advanced device structures. Accordingly, successful integration of such materials requires careful engineering and close attention to processing. In particular, silicates have been attractive since they combine some of the essential properties of silicon dioxide with larger values of the dielectric constant.