Model and Analysis of Gate Leakage Current
in Ultrathin Nitrided Oxide MOSFETs
(Advanced Topics in Semiconductor Devices)

Bin Li
Dept. of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A
November 27, 2002

Introduction

The paper is mainly discussing about Ultrathin Nitrided Oxide MOSFETs. It is necessary to use ultrathin gate to scale down of the size of the MOS devices. Nitridation treatment of thin oxides gate is used to improve the dielectric robustness to electric stress and to develop high permittivity gate dielectric stacks. This topic was recently discussed by Jonghwan Lee, Gijs Bosman, Keith R. Green, and D.Ladwig [1]. However, the reduction of the gate-oxide layer leads to increasing of the gate leakage current and the gate unreliabilities. So we have to find out appropriate methods to calculate the leakage current of the thin gate, meanwhile including the nitridation treatment. We discuss this topic in the following several steps. (A) First, we present a detailed study of a physics-based compact model for the gate leakage current of the ultrathin (less than 4.5 nm) gate oxide MOSFETs. Direct tunneling model (DT) is used to calculate the current density. (B) In the 2nd Part, we introduced a model called ITAT (Inelastic Trap-Assisted Tunneling), which is combined with a semi-empirical direct tunneling current to calculate the tunneling probability and tunneling current. Here, the nitridation treatment process is simulated by using trapping electronic states in the insulating oxide gate. (C) The 3rd part is related to the microscopic interpretation of the ITAT process. We consider the resonant tunneling through the oxide barrier containing potential wells corresponding to the defect states, and compare the current with the value obtained in the 2nd part. (D) In the 4th part, the current-voltage (I-V) and capacitance-voltage (C-V) simulation curves are given and compared with experimental results and those by previous classical models. The low-frequency noise results are also demonstrated. (E) In the last part, several comments and conclusions are made. We will try to address all these contents in detail.

Direct Tunneling Model of Ultrathin Gate Leakage Current

Usually, for the n-channel MOSFETs, we have three layers: n+ poly-Si, SiO₂, and p-Si at the inversion condition. Here, in order to make the discussion simpler, we use a different but similar structure: n+ poly-Si/SiO₂/n-Si, but where the n-Si substrate is in
accumulation condition, and it is actually a **n-channel** MOSFET. Since, we are dealing with ultra-thin oxide gate, so the difference in the potential energy over the oxide layer $qV_{ox}$ is usually smaller than the tunnel barrier height at the injecting interface of SiO$_2$/Si $\phi_b$. So we have to use direct tunneling model (DT) to calculate the tunneling current density [2] (Please refer to Fig. 1).

In Fig. 1, the case of electron tunneling from the accumulation layer in the Si substrate is considered with a tunnel barrier height $\phi_b$ and a positive applied gate bias voltage $V$ referent to the substrate. $E_{fs}$ and $E_{fp}$ are the Fermi energy levels in the Si-substrate and the poly-Si gate respectively. After the positive bias voltage $V$ applied to the n+ poly-Si/SiO$_2$/n-Si, a degenerate accumulation layer will occur in the Si substrate while a depletion layer is present in the poly-Si gate. We can use an independent electron approximation and an elastic tunneling process to get the tunneling current; while it is assumed that the transversal component of the electron energy $E_t$ is conserved during the tunneling in the MOS structure, $m_t$ is the transversal electron effective mass. So it gives the tunneling current density $J$ [3]:

$$J = \frac{4\pi \cdot qm_t}{h^3} \int \left[ E_{fs} - E \right] dE \int \left[ E - E_t \right] \cdot T_r(E, E_t) \ (1)$$

Where $E$ is the total energy of the tunneling electron measured from the Si conduction band (CB) edge and we can use the WKB approximation to calculate the tunneling transmission probability:

$$T_r(E, E_t) = \exp[-2\left(\frac{m_t}{h^2}\right)^{1/2} \int_0^x \sqrt{V(x) - E} \ dx]$$
For the tunneling case, $V(x) > E$, since $\sqrt{V(x) - E}$ is corresponding to the imaginary part of the wave vector in the oxide layer $\kappa_{ox}$, so the tunneling transmission probability becomes:

$$T_t(E, E_i) = T_t(E, k_i) = \exp[-2 \int_0^{x_1} \kappa_{ox}(E, k_i, x) \cdot dx] \quad (2)$$

with

$$\kappa_{ox} = k_i + i \kappa_{ox}$$

$$k_{ox}^2(x) = \frac{2 \cdot m_{ox}}{\hbar^2}[E - E_{c,ox}(x)]$$

$$k_i^2 = \frac{2 \cdot m_i E_i}{\hbar^2}$$

So we will get

$$T_t(E, E_i) = \exp[-\frac{2}{h} \int_0^{x_1} \sqrt{2m_i E_i - 2m_{ox}[E - E_{c,ox}(x)]} dx] \quad (3)$$

The energy of the SiO$_2$ Conduction Band edge $E_{c,ox}(x)$ is given by:

$$E_{c,ox}(x) = \phi_b + E_{F_S} - q F_{ox} \cdot x$$

And for the direct tunneling case $x_1$ is equal to the width of the oxide layer $t_{ox}$.

Replacing them into the equation (3), we have:

$$T_{DT}(E, E_i) = \exp[-\frac{2}{h} \int_0^{t_{ox}} \sqrt{2m_i E_i - 2m_{ox}(E + q \cdot F_{ox} \cdot x - \phi_b - E_{FS})} dx]$$

After integration, we will get our equation (4):

$$T_{DT}(E, E_i) = \exp[-\frac{4}{3} \left(\frac{2m_{ox}}{q\hbar}\right)^{1/2} \left(\phi_b + E_{FS} - E + \frac{m_i}{m_{ox}} E_i\right)^{3/2} - \left(\phi_b + E_{FS} - E + \frac{m_i}{m_{ox}} E_i - qV_{ox}\right)^{3/2} \cdot \frac{F_{ox}}{F_{ax}}] \quad (4)$$

where $V_{ax} = t_{ax} \cdot F_{ax}$

In Equation (4), if we define

$$f(E, E_i) = \left(\phi_b + E_{FS} - E + \left(\frac{m_i}{m_{ox}}\right) E_i\right)^{3/2} - \left(\phi_b + E_{FS} - E + \left(\frac{m_i}{m_{ox}}\right) E_i - qV_{ox}\right)^{3/2}$$
By using the first order approximation of Taylor Expansion around $E=E_{fs}$ and $E_t=0$, we have

$$f(E,E_t) = f(E,E_t)|_{E=E_{fs},E_t=0} + \frac{\partial f}{\partial E}|_{E=E_{fs},E_t=0} (E - E_{fs}) + \frac{\partial f}{\partial E_t}|_{E=0} E_t,$$

$$f(E,E_t) = \phi_b^{3/2} - (\phi_b - qV_{ox})^{3/2} + \frac{3}{2}[(\phi_b + \frac{m_t}{m_m} E_t - qV_{ox})^{1/2} - (\phi_b + \frac{m_t}{m_m} E_{fs})^{1/2}](E - E_{fs}) + \frac{3m_t}{2m_{ox}}[(\phi_b + E_{fs} - E)^{1/2} - (\phi_b + E_{fs} - E - qV_{ox})^{1/2}]E_t. \quad (5)$$

So now

$$T_{D}\ t(E,E_t) = \exp[-\frac{4}{3}(2m_{ox})\frac{1}{q\hbar}f(E,E_t)]$$

Using the double integration formula of $E$ and $E_t$ (Equation (1)), we will get the tunneling current:

$$J_D = \frac{A \cdot F_{ox}^2}{[1-(\phi_b - qV_{ox})^{1/2}]^2} \exp[-\frac{B}{F_{ox}} \frac{\phi_b^{3/2} - (\phi_b - qV_{ox})^{3/2}}{\phi_b^{3/2}}] \times \{1 - \exp[-\frac{3}{2} \frac{B}{F_{ox}} \frac{\phi_b^{1/2} - (\phi_b - qV_{ox})^{1/2}}{\phi_b^{1/2}} E_{fs}]\} \quad (6)$$

Where $F_{ox}$ is the average electric field cross the oxide layer, $F_{ox} = \frac{V_{ox}}{t_{ax}}$, $A$ and $B$ are two constants. When we considering the very strong degenerate accumulation in n-type Si-substrate, which is corresponding to the strong inversion condition at the p-type Si-substrate, we have $E_{fs} > \frac{4}{3} \frac{\phi_b^2}{q \cdot B \cdot t_{ox}}$. So the last term in the equation (6) is negligible small, we can get the simplified expression for the direct tunneling current:

$$J_D = \frac{A \cdot F_{ox}^2}{[1-(\phi_b - qV_{ox})^{1/2}]^2} \exp[-\frac{B}{F_{ox}} \frac{\phi_b^{3/2} - (\phi_b - qV_{ox})^{3/2}}{\phi_b^{3/2}}] \quad (7)$$

We have the parameters: $A = \frac{q^3}{16\pi^2\hbar\phi_b}$ and $B = \frac{4}{3}(2m_{ox})^{1/2} \frac{1}{q\hbar} \phi_b^{3/2}$. 
The voltage across the oxide layer $V_{ox}$ depends on the applied gate voltage $V_g$ as well as the surface potential $\varphi_s$ and the polysilicon drop $V_{poly}$ (We will discuss this equation a little bit more later, please see Equation (27)).

$$V_{ox} = V_g - V_{FB} - \varphi_s - V_{poly}$$  \hspace{1cm} (8)

By using some reference [4], we can get $V_{poly} = \frac{1}{2m} [2mV'_g + 1 - \sqrt{1 + 4mV'_g}]$

with

$$m = \frac{\varepsilon_{ox}^2}{2\varepsilon_a qN_{polys} t_{ax}^2} \quad \text{and} \quad V'_g = V_g - V_{FB} - \varphi_s.$$

Right now we have the direct tunneling current density $J_D$ ----- Equation (7), if we introduce a new model BSIM4 [5], we will get a new expression for the oxide tunneling current, which is similar to the equation (7), except for the correction function $C(V_g, t_{ox}, F_{ox}, \varphi_h)$.

$$C(V_g, t_{ox}, F_{ox}, \varphi_h) = \exp\left[\frac{20}{\varphi_h} \left(\frac{t_{ox}|F_{ox}| - \varphi_h}{\varphi_h} + 1\right)^\lambda \cdot \left(1 - \frac{t_{ox}|F_{ox}|}{\varphi_h}\right) \cdot \frac{V_g}{t_{ox}}\right]N$$

Using BSIM4 model, we can get the new gate oxide tunneling current:

$$J_{DT,BSIM4} = \frac{q^3}{8\pi h\varphi_h e_{ox}^3} \cdot C(V_g, t_{ox}, F_{ox}, \varphi_h) \cdot \exp\left[- \frac{8\pi \sqrt{2m_{ax} \varphi_h^3/2} [1 - (1 - |V_{ox}|/\varphi_h)^{3/2}]}{3hq|F_{ax}|}\right]$$  \hspace{1cm} (9)

We will use BSIM4 model in the following sections.

**Inelastic Trap-Assisted Tunneling (ITAT) Model**

In previous section, we introduce the BSIM4 model to represent the direct tunneling current of ultrathin MOS gate, now we will formulate an improved model to include the nitrided process of MOSFETs. This model combines the inelastic trap-assisted tunneling mechanism with the semi-empirical gate leakage current model of BSIM. In this process, the tunneling-in current is from the inversion layer to the traps and released to deeper positions; then the tunneling-out current is subsequently tunneling from the traps to the gate under the influence of the applied electric field. This two-step tunneling process includes a direct tunneling process into the trap sites with a tunneling probability $T_{DTi}$, and an additional direct tunneling step to the gate with a probability $T_{DTg}$. 
These two probabilities are related to the tunneling-in current $J_{DT,\text{in}}$ and the tunneling-out current $J_{DT,\text{out}}$ respectively (Please refer to Fig. 2.).

![Fig. 2](image)

Considering $x$ is the distance from the Si-SiO$_2$ interface, $N_{\text{trap}}(x, E)$ is the sheet trap density at $x$ with energy $E$, $f_t(x, E)$ is the electron occupancy function of the traps, $A_g$ is the tunneling area and $\sigma_i$ is the capture cross section in the traps.

Naturally, we can get the expression of tunneling-in and tunneling-out current as following:

\begin{align}
J_{DT,\text{in}} &= \frac{q}{A_g} \sigma_i N_{\text{trap}}(x, \phi_b - qF_{ox1} \cdot x) \cdot [1 - f_t(x, \phi_b - qF_{ox1} \cdot x)] g(\phi_b, x, F_{ox1}) \quad (10) \\
J_{DT,\text{out}} &= \frac{q}{A_g} \sigma_i N_{\text{trap}}(x, \phi_b - qF_{ox1} \cdot x) \cdot f_t(x, \phi_b - qF_{ox1} \cdot x) \cdot r(\phi_i, t_{ax} - x, F_{ox2}) \quad (11)
\end{align}

Where the tunneling-in electrons can only go into the un-occupied sites, while the tunneling-out electrons can only come out from the occupied sites, so we have factors $(1 - f_t)$ and $f_t$ in the equation (10) and equation (11) respectively. And the tunneling can be divided into two parts, $x$ and $t_{ax} - x$, the tunneling energy of the 2nd part should include the loss or decay energy in the trap, so we have $\phi_i = \phi_b - qF_{ox1} \cdot x + E_{loss}$. \[\]
So the final result of tunneling current of this inelastic trap assisted tunneling process is due to the detailed balance of \( J_{DT, in} \) and \( J_{DT, out} \) when the generation function \( g \) and recombination function \( r \) have the same expression. Since there are sequentially tunneling processes characterized by the tunneling rates \( g_{st} \) and \( r_{tg} \), so the total tunneling rate of a electron tunneling through the entire oxide via a trap site at location \( x \) in the oxide \( R \) should be satisfied the following equation:

\[
\frac{1}{R} = \frac{1}{g} + \frac{1}{r},
\]

so we have:

\[
R(E, x, F_{ox}) = \frac{g(\phi_b, x, F_{ox1}) \cdot r(\phi_b - qF_{ox2}x + E_{loss, t} - t_{ox} - x, F_{ox2})}{g(\phi_b, x, F_{ox1}) + r(\phi_b - qF_{ox1}x + E_{loss, t} - x, F_{ox2})} \quad (12)
\]

After obtaining the function \( R \), we can calculate the final tunneling current at the detailed balance condition, which is:

\[
J_{ITAT} (E, x, F_{ox}) = \frac{q}{A} \sigma \cdot N_{trap} (x, \phi_b - qF_{ox}x) \cdot R(E, x, F_{ox}) \quad (13)
\]

Where the local electric fields in expression (13), \( F_{ox1} \) and \( F_{ox2} \), charactering the separated tunneling regions can be obtained by using the equivalent MOS capacitor circuit and Gauss’s Law. Please see Fig.3.

![Fig.3](image)

So from \( \iiint_S \vec{E} \cdot d\vec{S} = \frac{1}{\varepsilon_{ox}} \iiint \rho \cdot dV \), we have \( (F_{ox1} - F_{ox}) \cdot \Delta S = \frac{Q_{trap}}{T_{ox}} (T_{ox} - x) \cdot \Delta S \).

Then we can get:

\[
F_{ox1} = F_{ox} + \frac{t_{ox} - x}{t_{ox}} \cdot \frac{qN_{trap}}{\varepsilon_{ox}} \quad (14)
\]

Using same algebra at the similar case, we have:
\[ F_{\text{ox2}} = F_{\text{ox}} - \frac{x}{l_{\text{ox}}} \cdot \frac{qN_{\text{trap}}}{\varepsilon_{\text{ox}}} \]  

(15)

Now we are going to calculate the tunneling rate function \( g \) or \( r \). Although they have a same expression under the detailed balance condition, when we implement them into Equation (13), we will use different parameters, so they will have different contributions to the final tunneling current \( J_{\text{ITAT}} \).

Very similar to the equation (1) \[ J = \frac{4\pi \cdot qm}{h^3} \int_0^E dE \int_0^{E_t} dE_t \cdot T(E,E_t) , \] which we used to calculate the direct tunneling current before, we have the uniform generation rate:

\[ g(\phi_b, x, F_{\text{ox1}}) = \frac{4\pi m_{\text{q}} A_{\text{g}}}{h^3} \int_0^E dE \int_0^{E_\perp} dE_\perp T_{\text{DT}} (\phi_b - E + E_{\text{FS}}, x, F_{\text{ox1}}, E_\perp) \]

Here, we can use the result of Equation (4) we obtained before

\[ T_{\text{DT}} (E, E_\perp) = \exp[-\frac{4}{3} \frac{(2m_{\text{ox}})^{1/2} (\phi_b + E_{\text{FS}} - E + m_{\text{ox}} E_{\perp})^{3/2} - (\phi_b + E_{\text{FS}} - E + m_{\text{ox}} E_{\perp} - qV_{\text{ox}})^{3/2}}{qh F_{\text{ox}}}] \]

We can just Let \( E \to \phi_b + E_{\text{FS}} - E \) and \( V_{\text{ox}} \to F_{\text{ox}} \cdot x \), we can get

\[ T_{\text{DT}} (E, E_\perp) = \exp[-\frac{4}{3} \frac{(2m_{\text{ox}})^{1/2} (E + m_{\text{ox}} E_{\perp})^{3/2} - (E + m_{\text{ox}} E_{\perp} - qV_{\text{ox}})^{3/2}}{qh F_{\text{ox}}}] \]  

(16)

At this moment, we can use 1\textsuperscript{st} order approximation of equation (16), then consider the BSIM4 model (in previous case the C function applied to the whole oxide gate region \( 0 \to t_{\text{ox}} \), but here it is only applied from the Si/SiO\textsubscript{2} interface to some specific trapping position \( x \)), so the whole calculation from now on will be identical with what we have done before. Finally, we will get a very similar expression as Equation (9). The only difference here is just the some constant multiplying factor:

\[ g(\phi_b, x, F_{\text{ox}}) = r(\phi_b, x, F_{\text{ox}}) = \frac{q^2}{16\pi^2 h_{\phi_b} \varepsilon_{\text{ox}}} \cdot C(x, F_{\text{ox}}, \phi_b) \cdot \exp[-\frac{4\sqrt{2m_{\text{ox}}} \phi_b^{3/2} [1 - (1 - x)F_{\text{ox}}/\phi_b]^{3/2}}{3hq F_{\text{ox}}}] \]

In order to get \( R(E, x, F_{\text{ox}}) \), we have to use \( g(\phi_b, x, F_{\text{ox}}) \) and \( r(\phi_r, t_{\text{ox}} - x, F_{\text{ox2}}) \).

So we can define several constants:
\[ A_i = \frac{q^2 A_g}{16\pi^2 \hbar e_{ox}} \quad B_i = \frac{4\sqrt{2m_{ox}}}{3q\hbar} \]
\[ C(x, F_{ox}, \phi_b) = \exp[\frac{20}{\phi_b} (\frac{x|F_{ox}| - \phi_b}{\phi_{bo}} + 1)^2 \cdot (1 - \frac{t_{ax}|F_{ax}|}{\phi_b})] \cdot F_{ax} \cdot N \]

Now we can write down:
\[ g(\phi_b, x, F_{ax1}) = A_i \cdot \frac{C(\phi_b, x, F_{ax1})}{\phi_b} \cdot \exp[-B_i \cdot \frac{\phi_b^{3/2}}{F_{ax1}} \cdot \beta(\phi_b, x, F_{ax1})] \quad (17) \]
\[ g(\phi_t, t_{ax} - x, F_{ax2}) = A_i \cdot \frac{C(\phi_t, t_{ax} - x, F_{ax2})}{\phi_t} \cdot \exp[-B_i \cdot \frac{\phi_t^{3/2}}{F_{ax2}} \cdot \beta(\phi_t, t_{ax} - x, F_{ax2})] \quad (18) \]

Where we have \( \beta(\phi_b, x, F_{ax1}) = 1 - (1 - \gamma_{t1} F_{ax1})^{3/2} \) with \( \gamma_{t1} = \frac{q^2 x}{\phi_b} \),
\[ \beta(\phi_t, t_{ax} - x, F_{ax2}) = 1 - (1 - \gamma_{t2} F_{ax2})^{3/2} \) with \( \gamma_{t2} = \frac{q(t_{ax} - x)}{\phi_t} \).

Then we insert the equation (17) & (18) we obtained here into the equation (12), (13), we will get the tunneling current density of ITAT model.
\[ J_{ITAT}(E, x, F_{ax}) = \frac{q}{A_g} \sigma \cdot N_{trap}(x, \phi_b - q F_{ax1}) \cdot R(E, x, F_{ax}) \]
\[ = \frac{q}{A_g} \sigma \cdot N_{trap}(x, \phi_b - q F_{ax1}) \cdot \frac{g(\phi_b, x, F_{ax1})}{1 + g(\phi_b, x, F_{ax1}) / \rho(\phi_b - q F_{ax1} x + E_{loss}, t_{ax} - x, F_{ax2})} \]

After several steps’ calculations, we can get the following expression straightly.
\[ J_{ITAT}(\phi_b, \phi_t, x, F_{ax}) = \frac{q A_i}{A_g} \sigma N_{trap}(x, \phi_b - q F_{ax1}) \]
\[ \times \frac{C(\phi_b, x, F_{ax1}) \cdot \exp[-B_i \cdot \frac{\phi_b^{3/2}}{F_{ax1}} \cdot \beta(\phi_b, x, F_{ax1})]}{1 + \frac{\phi_t}{\phi_b} \cdot C(\phi_t, t_{ax} - x, F_{ax2}) \cdot \exp[-B_i \cdot \frac{\phi_t^{3/2}}{F_{ax2}} \cdot \beta(\phi_t, t_{ax} - x, F_{ax2})]} \quad (19) \]

It is quite complicated!
Microscopic Interpretation of Resonant Current Transport

For this part, the original paper mainly discussed the resonant tunneling through the oxide barrier containing potential well associated with quantum electronic states. In this model, the transitions among the local bounded electron states in the trap are interpreted as the process of the interactions of the electrons with phonons and photons. The resonance tunneling process is non-negligible since it causes an enhancement in the tunneling current density when the trap energy level and electron energy level line up. And the tunneling processes suffer inelastic events, so it should be modified to include the inelastic scattering processes occurring in the trapping state.

Unfortunately, I couldn’t find the references related to this part, so I can only give a brief introduction and discussion about this complicated theory here.

By using the formula for the Tsu-Esaki tunneling current of the single resonant energy level, we have [6],

\[ J = \frac{e}{2\hbar \pi^2} \int dk' k' |T|^2 (E_F - E) \]

Where \( k_F(E_F) \) is the Fermi wave vector (Fermi energy), \( k' (k') \) is the magnitude of the electron wave vector in the left emitter (right collector), \( T \) is the left to right transmission probability.

Since right now we are discussing about a system with \( N_{res} \) different tunneling resonance energy levels in the trap, so we can extend this formula to include the contributions from all.

\[ J_{RES}(E_{res}^{res}, x, F_{ox1}, F_{ox2}, p) = \frac{qm_{ox}}{2\hbar \pi^2} \sum_{n=1}^{N_{res}} (E_F - E) \cdot T_n^{tot}(E, x, F_{ox1}, F_{ox2}, p) dE \] (20)

Which is given as equation (19) in the original paper, where \( T_n^{tot}(E, x, F_{ox1}, F_{ox2}, p) \) ---- the total tunneling probability for an electron to traverse through a double barrier with a quantum well is given as

\[ T_n^{tot}(E, x, F_{ox1}, F_{ox2}, p) = \sum_n T_n^{peak}(E, x, F_{ox1}, F_{ox2}) \cdot \frac{\Gamma_n(E, x, F_{ox1}, F_{ox2}, p) \Gamma_n(E, x, F_{ox1}, F_{ox2})}{(E - E_n^{res})^2 + \Gamma_n^2(E, x, F_{ox1}, F_{ox2}, p)} \] (21)

Where I have a different idea of the peak value at the \( nth \) resonance energy \( T_n^{peak}(E, x, F_{ox1}, F_{ox2}) \) with the authors. Since there is a two-step sequential tunneling
process from emitters to trap states (from 0 to $x$), then from trap states to collectors (from $x$ to $t_{ox}$), so it is better if the formula is:

$$T_{n}^{\text{peak}}(E, x, F_{\text{ox}1}, F_{\text{ox}2}) = \frac{4 \cdot T_{\text{DT,st}}(E, x, F_{\text{ox}1}) \cdot T_{\text{DF,lg}}(E, t_{ox} - x, F_{\text{ox}2})}{[T_{\text{DT,st}}(E, x, F_{\text{ox}1}) + T_{\text{DT,lg}}(E, t_{ox} - x, F_{\text{ox}2})]^2}$$

which is different with the equation (18a) in the original paper.

By using the equation (18b), (18c), (18d) and (18e) in the original paper, we can get the following result:

$$\Gamma_{n}(E, x, F_{\text{ox}1}, F_{\text{ox}2}, p) = \frac{\Gamma_{n}^{e}(E, x, F_{\text{ox}1}, F_{\text{ox}2}) \cdot E_{n}^{\text{res}}(E, x, F_{\text{ox}1}, F_{\text{ox}2})}{(E - E_{n}^{\text{res}})^2 + \Gamma_{n}^{e}(E, x, F_{\text{ox}1}, F_{\text{ox}2}) \cdot \Gamma_{n}^{i}(v, p)}$$

$$= \frac{\Gamma_{n}^{e}(E, x, F_{\text{ox}1}, F_{\text{ox}2}) + \Gamma_{n}^{e}(E, x, F_{\text{ox}1}, F_{\text{ox}2}) \cdot \Gamma_{n}^{i}(v, p)}{(E - E_{n}^{\text{res}})^2 + \Gamma_{n}^{i}(v, p)^2 + 2 \Gamma_{n}^{i}(E, x, F_{\text{ox}1}, F_{\text{ox}2}) \cdot \Gamma_{n}^{i}(v, p)}$$

$$= \frac{(T_{\text{DT,st}} + T_{\text{DT,lg}})^2 + 2 p \cdot (T_{\text{DT,st}} + T_{\text{DT,lg}})}{(T_{\text{DT,st}} + T_{\text{DT,lg}} + 2 p)^2 + \frac{(E - E_{n}^{\text{res}})}{\hbar v_{n}})^2}$$

So the total resonant tunneling probability can be expressed as:

$$T_{n}^{\text{tot}}(E, x, F_{\text{ox}1}, F_{\text{ox}2}, p) = \sum_{n} \frac{4 \cdot T_{\text{DT,st}} \cdot T_{\text{DT,lg}}}{(T_{\text{DT,st}} + T_{\text{DT,lg}})^2 + 2 p \cdot (T_{\text{DT,st}} + T_{\text{DT,lg}})}$$

From the equation (4)

$$T_{\text{DT}}(E, E_{\perp}) = \exp[-\frac{4 (2m_{\text{ox}})^{1/2} (\phi_{b} + E_{FS} - E + \frac{m_{l}}{m_{\text{ox}}} E_{\perp})^{3/2} - (\phi_{b} + E_{FS} - E + \frac{m_{l}}{m_{\text{ox}}} E_{\perp} - qV_{\text{ox}})^{3/2}}{3 q h} F_{\text{ox}}]$$

We can estimate the expressions for $T_{\text{DT,st}}$ and $T_{\text{DT,lg}}$ respectively:

$$T_{\text{DT,st}} = \exp[-\frac{4 (2m_{\text{ox}})^{1/2} (\phi_{b} + E_{FS} - E + \frac{m_{l}}{m_{\text{ox}}} E_{\perp})^{3/2} - (\phi_{b} + E_{FS} - E + \frac{m_{l}}{m_{\text{ox}}} E_{\perp} - qF_{\text{ox}1} \cdot x)^{3/2}}{3 q h} F_{\text{ox}1}]$$
Just let \( \phi_b \rightarrow \phi_t = \phi_b - qF_{ax1} \cdot x + E_{loss} \cdot x \rightarrow t_{ax} - x \) then

\[
T_{DT,ig} = \exp\left[ -\frac{4}{3} \frac{(2m_{ax})^{1/2}}{q\hbar} \left( \phi_t + E_{FS} - E + \frac{m_{ax}}{m_{ax}^2} E_{\perp} \right)^{3/2} - (\phi_t + E_{FS} - E + \frac{m_{ax}}{m_{ax}^2} E_{\perp} - qF_{ax2} \cdot (t_{ax} - x)^{3/2} \right]
\]

If we apply these two terms into the equation (22), after that put the equation (22) into the integration (20), from there we will see a very complicated expression. I have tried to calculate it for quite a while, but I failed to get any reasonable result, it might be required to use some special magic algebra to get it. Here, I just write down the result:

\[
J_{RES}(E_n^{res}, x, F_{ax1}, F_{ax2}, p) = \frac{q \cdot m_{ax}}{2\pi^2 \hbar^3} \sum_{n=1}^{N_{ax}} \frac{\Gamma_n^{DT, st} \Gamma_{n}^{DT, sg}}{\Gamma_n^{DT, st} + \Gamma_n^{DT, sg}} \times \{2(E_{fs} - E_n^{res})[\tan^{-1}\left(\frac{2(E_{fs} - E_n^{res})}{\Gamma_n}\right)] - \frac{\Gamma_n}{2} \ln\left[\frac{(E_{fs} - E_n^{res})^2 + \Gamma_n^2/4}{(E_n^{res})^2 + \Gamma_n^2/4}\right]\}
\]

(23)

When we are considering it in the resonant region, where \((E_{fs} - E_n^{res}) \gg \Gamma_n\) and \(E_n^{res} \gg \Gamma_n\), we will have

\[
\tan^{-1}\left(\frac{2(E_{fs} - E_n^{res})}{\Gamma_n}\right) \approx \tan^{-1}(\infty) = \frac{\pi}{2} \quad \text{and} \quad \tan^{-1}\left(\frac{2E_n^{res}}{\Gamma_n}\right) \approx \tan^{-1}(\infty) = \frac{\pi}{2}
\]

Meanwhile, the last term \(\frac{\Gamma_n}{2} \ln\left[\frac{(E_{fs} - E_n^{res})^2 + \Gamma_n^2/4}{(E_n^{res})^2 + \Gamma_n^2/4}\right]\) is negligible since \(\Gamma_n \ll E_{fs} - E_n^{res}\).

So we can get

\[
J_{RES}(E_n^{res}, x, F_{ax1}, F_{ax2}, p) = \frac{q \cdot m_{ax}}{\pi \cdot \hbar^3} \sum_{n=1}^{N_{ax}} \frac{\Gamma_n^{DT, st} \Gamma_{n}^{DT, sg}}{\Gamma_n^{DT, st} + \Gamma_n^{DT, sg}} \times 2(E_{fs} - E_n^{res})[\frac{\pi}{2} + \frac{\pi}{2}] = \frac{q \cdot m_{ax}}{\pi \cdot \hbar^3} \sum_{n=1}^{N_{ax}} \frac{\Gamma_n^{DT, st} \Gamma_{n}^{DT, sg}}{\Gamma_n^{DT, st} + \Gamma_n^{DT, sg}} (E_{fs} - E_n^{res})
\]

(24)

So at this spot, we derived the equation (21) in the original paper, it shows that the sum of the resonant tunneling current components in the different resonant energy levels will give the total resonant current density here, and it is complementary to the ITAT current density which we discussed previously (in Equation (19)).
Comparison between modeling simulation and experiment results

At the beginning of this part, the original paper addressed the fabrication of Nitrided Ultra thin Oxides gate, but we are going to neglect it since it is not an important problem all over this paper. Then it shows a schematic energy band diagram by a special program simulation (Fig.2 in Jonghwan Lee et al’s paper), here we can see when we use a large value of gate voltage and high implanted charge concentration in the p-substrate, the band-bending is large, the energy level of conduction band edge at the SiO$_2$/Si interface is even lower than the Fermi level, which is corresponding to a very strong inversion case. Also we can see the energy distribution in the potential well, the ground state and the 1$^{\text{st}}$ excited state are dominated. We illustrate this diagram as Fig. 1 in our appendix page.

Now we are going to discuss the C-V characteristics of this kind of ultra-thin gate. As the total capacitance of the gate is in seriated combination of three parts: silicon capacitance $C_{si}$, oxide capacitance $C_{ox}$, and the capacitance due to the polysilicon depletion region $C_{poly}$, we get the total capacitance $C_g = C_{total}$:

\[
\frac{1}{C_{total}} = \frac{1}{C_{ox}} + \frac{1}{C_{poly}} + \frac{1}{C_{si}} \quad (25)
\]

Where $C_{ox} = \varepsilon_{ox} t_{ox}$ and $C_{poly} = \frac{dQ_{poly}}{dV_{poly}} = -\frac{dQ_s}{dV_{poly}}$.

Since the induced positive charge in the polysilicon depletion region $Q_{poly}$ is equal to the space charge density in the silicon inversion region $Q_s$, but with an negative sign.

We can easily get $\frac{dQ_{poly}}{dV_{poly}} = \frac{q\varepsilon_{si} N_{poly}}{Q_{poly}} = \frac{q\varepsilon_{si} N_{poly}}{|Q_s|}$, where $N_{poly}$ is doping concentration in the poly-silicon region (cm$^{-3}$), but the $Q_{poly}$ is the sheet charge density of the polysilicon in the depletion region (q·cm$^{-2}$), so it will give us the correct dimension [7].

Now we are going to discuss the most complicated part $C_{si}$, the channel capacitance. We can consider this capacitance in two parts in series: one is the contribution from the electron dynamic density of states near the Fermi level $C_{dos}$, the other is the capacitance of the electron in the inversion layer $C_{inv}$ (considering the average position of electrons in the inversion layer referenced to the SiO$_2$/Si interface is $z_{av}$). So naturally, we will have the following formula.

\[
C_{si} = q \frac{\partial N_{str}}{\partial \psi_x} = (C_{dos}^{-1} + C_{inv}^{-1})^{-1} = \left(\frac{dV_F}{dQ_F} + \frac{\gamma \cdot z_{av}}{\varepsilon_{si}}\right)^{-1}
\]
Here, we introduced a parameter $\gamma$, we will talk more about it later. The total carriers’ density is $N_{tot} = \frac{Q}{q}$.

Near the Fermi level, we have $\frac{dV_F}{dQ_F} = \frac{d(E_F / q)}{d(q \cdot n)} = \frac{1}{q^2 (dn / dE_F)}$.

Where $n$ is the density of state, and at the room temperature we get $\frac{dn}{dE_F} = \frac{N_{tot}}{kT} = \frac{|Q_s|}{qkT}$, so $C_{des}^{-1} = \frac{dV_F}{dQ_F} = \frac{kT}{q|Q_s(\psi_s)|}$.

From some related reference, we have the expression of $z_{av}$, since the discussion is several pages long [8], so here we just use the result directly.

$z_{av} = 3(4\pi^2 m^* q^2 / \epsilon_0 h^2)^{-1/3} \cdot (N_{dep} + (11/32)|Q_s(\psi_s)| / q)^{-1/3}$

Where constant $N_{dep}$ is the depletion charge density, we have $N_{dep} = qN_A \cdot W$, $W = \sqrt{\frac{2\epsilon_{si}\psi_s}{qN_A}}$, so $N_{dep} = \sqrt{2q\epsilon_{si}N_A\psi_s}$.

Up to this point, we have got the expressions of those three capacitances respectively, where $C_{ox}$ is a constant, $C_{poly}$ and $C_{ai}$ are both functions of $Q_s$; that’s to say, they are both the functions of surface potential $\psi_s$. Since as we know $Q_s(\psi_s) = \sqrt{2\epsilon_{si} kT} F\left(\frac{q\psi_s}{kT}\right) - qN_{dep}$.

Here, the definition of $Q_s$ is deferent with what have learned from class [9] by the term $q \cdot N_{dep}$, the $Q_s$ here just includes the inversion charge, exclusion of the depletion charge due to the static negative ions in the p-substrate. I guess that when the author use capacitance meter to measure the capacitance of the whole system, they just measure from the polysilicon gate to SiO$_2$/Si interface, so without considering the depletion charge seems OK to me.

$L_D$ is the Debye length, which is defined by $L_D = \sqrt{\frac{\epsilon_{si} kT}{q^2 N_A}}$.

Function $F\left(\frac{q\psi_s}{kT}\right) = \sqrt{\frac{1}{[\exp(-\frac{q\psi_s}{kT}) + \frac{q\psi_s}{kT} - 1]} + \left(\frac{n_i}{N_A}\right)^2 \cdot \frac{\exp\left(\frac{q\psi_s}{kT}\right) - \frac{q\psi_s}{kT} - 1]}{kT} - \frac{q\psi_s}{kT} - 1)}$

Since $N_A = n_i e^{q\psi_s / kT}$ and $n_{po} = n_i \exp(-q\psi_B / kT)$, so $\frac{n_{po}}{N_A} = \left(\frac{n_i}{N_A}\right)^2$, which shows that F function here is identical to the expression that we got in our class.
Now we can give the expression of the gate capacitance of the nitrided ultra-thin oxide MOSFETs in terms of $Q_s$ (or directly $\psi_s$).

\[
\frac{1}{C_{\text{total}}} = \frac{t_{\text{ox,eq}}}{\varepsilon_{\text{ox}}} + \left| \frac{Q_s(\psi_s)}{q\varepsilon_{\text{si}}N_{\text{poly}}} \right| + \frac{kT}{q\varepsilon_{\text{si}}} \left( \frac{48\pi^2 m^* q^2}{\varepsilon_{\text{si}} h^2} \right)^{-1/3} \cdot \left( \sqrt{2q\varepsilon_{\text{si}} N_{\text{A}}\psi_s} + \frac{11}{32} \frac{Q_s(\psi_s)}{q} \right)^{-1/3}
\]

(26)

Where $m^*$ is the effective electron mass and $\gamma$ is introduced here to make a correction of the effective width of the inversion layer. $\gamma$ depends weakly on the value of $\frac{Q_s(\psi_s)}{q}$, after the empirical estimation and practical self-consistent simulation, they obtained an expression for Si (100) surface: $\gamma = 0.41 + \exp\left(-\frac{\left| \frac{Q_s(\psi_s)}{qN_{\text{m}}} \right|}{10}\right)$, constant $N_{\text{m}}$ is given as a value of $1.264 \times 10^{11} \text{cm}^{-2}$ [10].

In the notes of our course, we have the formula for gate in terms of surface potential:

\[
V_g = V_{FB} + V_{SB} + \psi_s + \frac{|Q_s|}{C_{\text{ox}}} = V_{FB} + V_{SB} + \psi_s - \frac{Q_s}{C_{\text{ox}}}
\]

But here $Q_s$ we used here is only the inversion charge without including the depletion charge, so we’ve got to modify this formula. Actually, we just need to add an additional term $V_{\text{poly}}$ (the voltage drop in the polysilicon depletion region, which is negative) on the right side of the equation, it is still true and more accurate. Now we assume substrate grounded so $V_{SB} = 0$, so we can have

\[
V_g = V_{FB} + V_{\text{poly}} + \psi_s - \frac{Q_s}{C_{\text{ox}}} \quad (27)
\]

$V_{FB}$ flat-band voltage depends on energy barrier potential of the Si-substrate side, which is easy to calculate or measure; we can also measure $V_{\text{poly}}$ without too much trouble. Then by using Equation (27), we can solve the equation for the surface potential $\psi_s$ as a function of gate voltage $V_g$. Then just put $\psi_s(V_g)$ in the equation (26), we can get expression $C_{g_s}$ (same as $C_{\text{total}}$ ) in term of $V_g$, and plot a curve “$C_g$ vs.$V_g$”.

In the original pager, Fig.3 gives the simulation result of this model, comparing with classical model and experimental result. From the plot, we see the current model
works much better than the previous classical model. (We also attached this plot in the appendix as Fig. 2.) One thing I have to mention is that when we use $t_{ox,eq}$ to do calculation in equation (26), we should know our gate oxide is nitrided, so $t_{ox,eq}$ is different with the measured value $t_{SION}$, we have

$$t_{ox} = \frac{\varepsilon_{ox}}{\varepsilon_{SION}} t_{SION} = \frac{3.9\varepsilon_0}{5.7\varepsilon_0} t_{SION}.$$  

Although the difference between the simulation result by the present model and experiment result is very small, it obviously exists, it might be due to the interface trapping charge. Now we want to evaluate the interface trap density from the quasi-static C-V technique. And we will try to separate the interface trap capacitance $C_{it}$ from the silicon capacitance. From equation (25), we know for the measured total gate capacitance $C_{total,m}$:

$$\frac{1}{C_{total,m}} = \frac{1}{C_{ox}} + \frac{1}{C_{poly}} + \frac{1}{C_{si} + C_{it}}$$

So we have

$$C_{it} = \left( \frac{1}{C_{total,m}} - \frac{1}{C_{ox}} - \frac{1}{C_{poly}} \right)^{-1} - C_{si}$$

From the previous discussion, we have got

$$\frac{1}{C_{ox}} = \frac{t_{ox,eq}}{\varepsilon_{SION}} = \frac{\varepsilon_{ox}}{\varepsilon_{SION}} t_{SION}, \quad \frac{1}{C_{poly}} = \frac{3.9\varepsilon_0}{5.7\varepsilon_0} t_{SION}, \quad \frac{1}{C_{si}} = \frac{|Q_s(\psi_s)|}{q\varepsilon_{si}N_{poly}}$$

$$C_{si} = \left\{ \frac{kT}{q|Q_s(\psi_s)|} + \frac{3}{\varepsilon_{si}} \left[ 0.41 + \exp\left[ -\left( \frac{|Q_s(\psi_s)|}{qN_m} \right)^{1/2} \right] \right] \right\} \left( \frac{48\pi^2m^*q^2}{\varepsilon_{si}h^2} \right)^{-1/3} \cdot \left( \sqrt{2q\varepsilon_{si}N_s} \psi_s + \frac{11}{32} \frac{|Q_s(\psi_s)|}{q} \right)^{-1/3}$$

Since $C_{total,m}$ is a practically measurable value, so we find that $C_{it}$ is actually a function of $\psi_s$.

Now if we define the interface trap charge density by $N_{it} = \frac{C_{it}}{q}$, we will get

$$N_{it}(\psi_s) = \frac{C_{it}(\psi_s)}{q} = \frac{1}{q} \left[ \frac{C_{total,m}C_{ox}C_{poly}}{C_{total,m}C_{poly} - C_{total,m}C_{poly} - C_{total,m}C_{ox}} - C_{si}(\psi_s) \right] \quad (28)$$

Just use this formula, we can get the plot “$N_{it}$ vs. $\psi_s$” (Please see Fig.4 in the original paper by Jonghwan Lee et al), which give us the comparison between the
simulation result of this quantum related model and that of the classical model, where we can see results from those two models are comparable, but the present model is more smooth. In the region near the band edge, we can only use this quantum model to get the interface trap density. We also give this picture as Fig. 3 in appendix.

Now let’s switch into the discussion of “I-V” characteristics.

In previous discussion, we have got the equation (19)

\[
J_{ITAT}(\phi_b, \phi_i, x, F_{\text{ox}}) = \frac{q A}{A_g} \sigma, N_{\text{trap}}(x, \phi_b - qF_{\text{ox}} x) \\
\times \frac{C(\phi_b, x, F_{\text{ox}})}{\phi_b} \cdot \exp[-B_1 \cdot \frac{\phi_b^{3/2}}{F_{\text{ox}}} \cdot \beta(\phi_b, x, F_{\text{ox}})] \\
+ \frac{C(\phi_b, x, F_{\text{ox}})}{\phi_b} \cdot \exp[-B_1 \cdot \frac{\phi_b^{3/2}}{F_{\text{ox}}} \cdot \beta(\phi_b, x, F_{\text{ox}}) - \frac{\phi_i^{3/2}}{F_{\text{ox}}} \beta(\phi_i, x - x, F_{\text{ox}})]
\]

Where we can see the trapping density \(N_{\text{trap}}\) is a position and energy dependent quantity, but here we assume it is a constant: \(N_{\text{trap}} = N_{\text{tr}}\) by using equation (28) at a certain value of surface potential \(\psi_s\). Since \(N_{\text{tr}}(\psi_s)\) and \(\psi_s(V_g)\), so \(N_{\text{trap}}\) is only a function of gate voltage \(V_g\) under this consideration.

If we know the energy barrier height \(\phi_b\), measured the average electrical field in the nitrided oxide layer \(F_{\text{ox}} = \frac{V_{\text{ox}}}{t_{\text{ox},eq}}\), put some reasonable values for the trap capture cross section \(\sigma\), and choose a reasonable value for the constant factor \(\lambda\) in the C function, when we are considering it at a fixed value of \(V_g\), the equation (19) will be just a function of position \(x\). On the other hand, we can choose a fixed point in the oxide layer to simulate this formula, then we will get the tunneling current density of the ITAT model is just a function of gate voltage \(V_g\): \(J_{ITAT}(V_g)\).

The original paper gave the ITAT modeling result at certain condition for \(n + poly-Si - nMOSFET\), they used the following parameters:

\[
t_{\text{SON}} = 3.21\text{nm} \quad t_{\text{ox},eq} = 2.2\text{nm} \quad \phi_{b,\text{SON}} = 2.6eV \quad \phi_i = 2.5eV \\
m_{\text{ox}} = 0.4m_0 \quad \sigma_i = 5.0 \times 10^{-14} \text{cm}^2 \quad x = x_i = 0.46t_{\text{SON}} \quad \text{(trap position)} \\
\lambda = 0.6 \quad \text{and trapping density used in function C: } N_{\text{trap}} = 2 \times 10^{-2} \text{cm}^2.
\]

From the Fig. 5(a) of the original paper (Fig. 4 in our appendix), we can see after implementation of these pre-defined parameters, the ITAT model works very well. It
almost fits every point of the experimental data, while the DT model (direct tunneling) seems deviated away from the experiment result, which shows the advantage of using ITAT model in “I-V” characteristics. This model is also equally applicable to p-channel MOSFET, just modified the parameters a little bit, we will see the similar result.

If we keep the gate voltage $V_g$ and energy barrier $\phi_s$ constant in the equation (19), while continuously change the trap position $x_t$ and trap energy $\phi_t$ in some range, so the current density will become as a two dimensional function $J_{ITAT}(x_t, \phi_t)$. In the Fig. 6(a) and Fig.6(b) illustrate the “3-D plot of influence of trap energy levels and trap positions on current density” and “contour plot of this 3-D graph at certain current densities”. We can easily see that the current peaks for traps located in the middle of the oxide layer and decay quickly for trap positions away from the trapping center and at deeper energy levels (You can see this in the Fig. 5 of the appendix).

We are not going to say a lot about the noise of leakage current, since I do not quite understand the technology named remote plasma nitried oxide (RPNO) process, which is involved in the measurement here. However, obviously, we can see the $1/f$ noise from the Fig. 8 of the original paper, which indicates the existence of inelastic process occurring in the trap states and the slow relaxation times in the oxide barrier.

Discussion

After putting lots of effort in studying this topic, I get the main idea and see the advantage of applying inelastic trap-assisted tunneling (ITAT) model combined with the semi-empirical direct tunneling current (DT) model and (BSIM4) model to simulate gate leakage current in ultrathin nitrided oxide MOSFETs. By analyzing of the “C-V” and “I-V” characteristics, we can see that the ITAT model can get much better simulation results compared with the previous classical models. One achievement here is to consider this problem as a two-step inelastic-assisted tunneling process through the whole gate oxide, using WKB approximation to calculate the generation and recombination rates in the trapping well, and finally the tunneling current. Right now, I still have trouble to understand and derive several formulas completely. However, I have obtained lots of fruit after I finished writing this paper, and have ever been able to consider the potential application of ultrathin Nitrided Oxide MOSFETs.
Reference


Appendix

Fig. 1 Schematic band diagram of an $n^+$ polysilicon-SiO$_2$-p-Si MOS structure

Fig. 2 Comparison of C-V characteristics by the proposed model and the classical model, and measured data for $n^+$ polysilicon-Si nMOSFET
Fig. 3 Evaluation of interface trap density simulated by the proposed model and the classical model for $n^+$ polysilicon-Si nMOSFET

Fig. 4 Comparison of I-V characteristics simulated by the proposed model and DT model, and measured data
Fig. 5 Influence of trap energy levels and trap positions on the tunneling current density for n⁺ polysilicon-Si nMOSFET