Leakage Current in Deep-Submicron CMOS

Renesas Design Vietnam Co., Ltd.
CDG Group - Analog2 team
Date: July 10th 2014

Name : Vinh Le
Mail: vinh.le.wh@rvc.renesas.com

Outline

0. Introduction
1. Sources of leakage current
   + pn Junction Reverse-Bias Current
     > Band-to-Band Tunneling Current
   + Subthreshold Leakage
     > Weak inversion
     > Drain-Induced Barrier Lowering
     > Body Effect
     > Narrow-Width Effect
     > Effect of channel Length & Vth Rolloff
     > Effect of Temperature
   + Tunneling into and Through Gate Oxide
     > Fowler–Nordheim Tunneling
     > Direct Tunneling

Outline

1. Sources of leakage current(cont.)
   + Injection of Hot Carriers from Substrate to Gate Oxide
   + Gate-Induced Drain Leakage
   + Punchthrough
2. Leakage reduction technique
   + Channel Engineering for Leakage Reduction
     > Retrograde Doping
     > Halo Doping
   + Circuit Techniques for Leakage Reduction(optional)
     > Standby Leakage Control Using Transistor Stacks (Self-Reverse Bias)
     > Multi V_{th} design
     > Dynamic V_{th} design
3. Conclusions

Introduction

Introduction

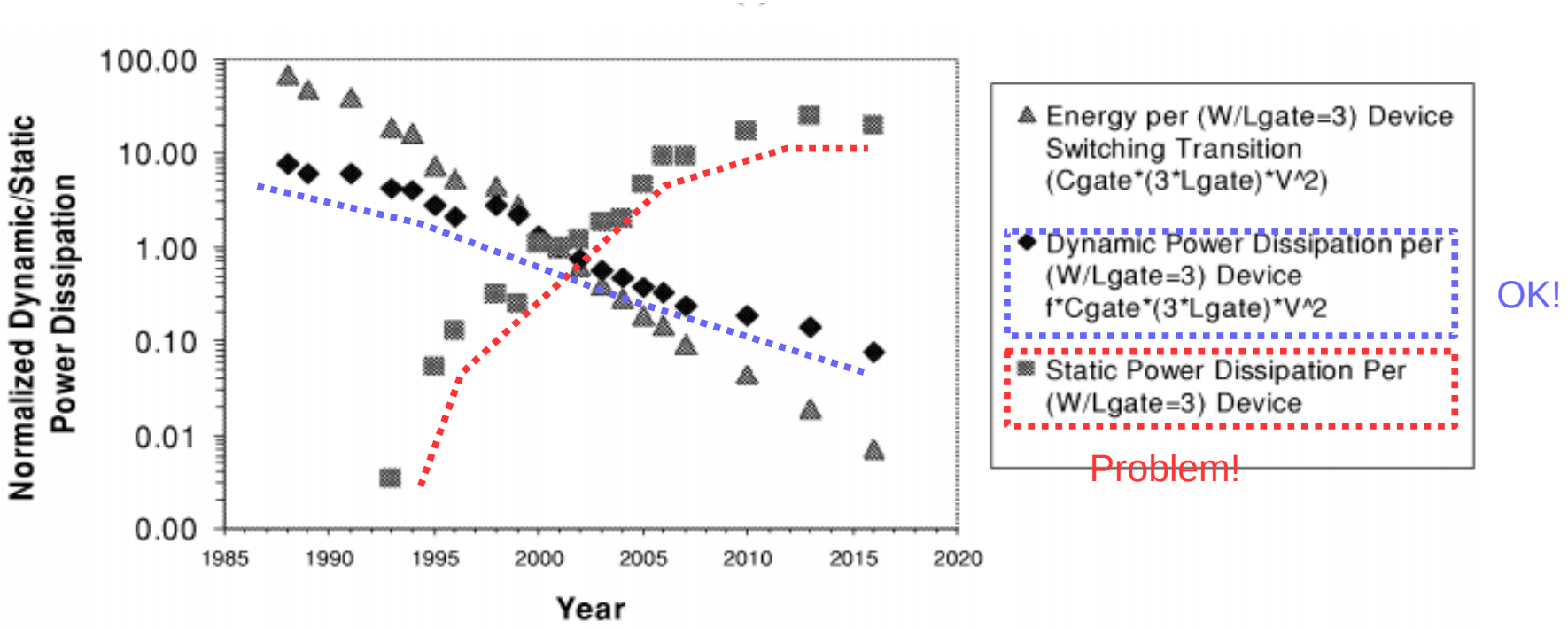
- Power density in high performance VLSI chips is already at the same level as in a nuclear reaction.

- Battery maximum power and capacity increase by 10-15% per year, but chip power requirements increase much faster: 35-40% per year.

Introduction

- VDD has been scaled down in order to keep power consumption under-control => V_{th} has to be scaled to maintain performance.

- As V_{th} scaled, the substantial increase of the sub-threshold leakage current.



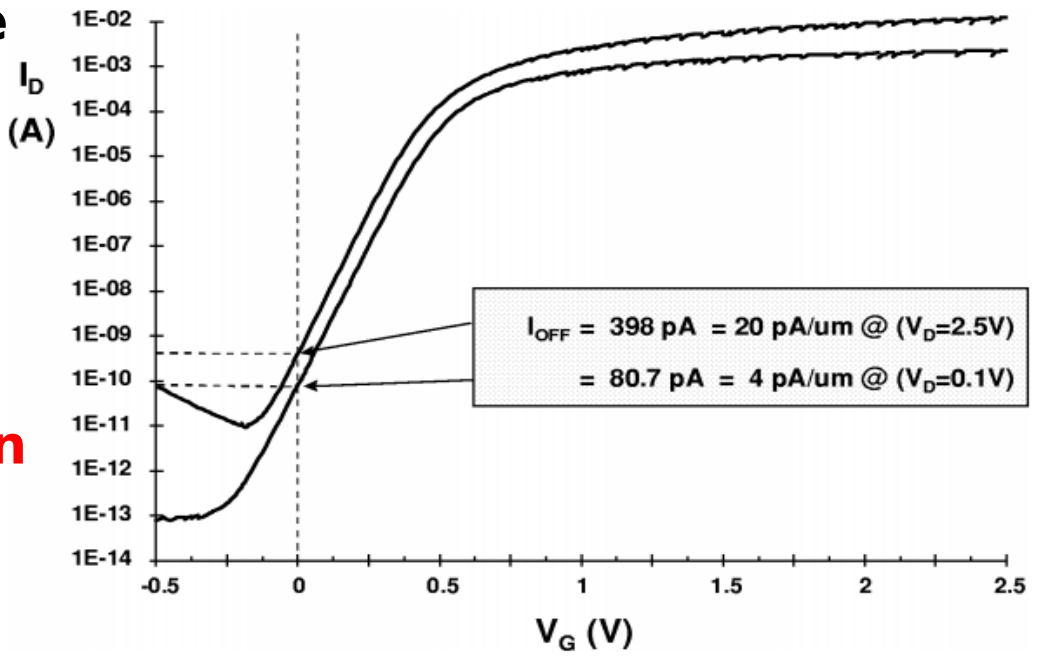
Introduction

- Transistor off-state current (I_{OFF}) is the drain current when the gate-source voltage is zero.

- The RHS(*) graph shows the example of off-current of n-channel transistor.

- I_{OFF} is influenced by:

Threshold Voltage, dimension,channel doping profile, drain/source junction depth, gate oxide thickness and VDD.



The n-channel transistor drain current

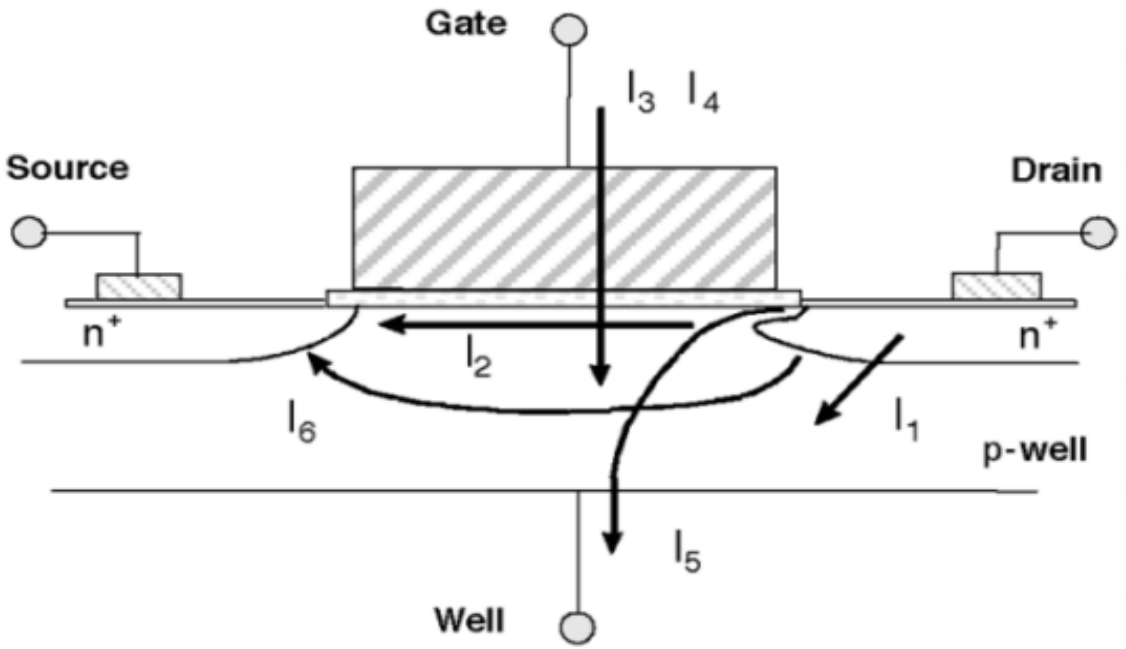
(*)RHS : right hand side

Introduction

- In short channel devices, the reduction V_{th} causing substantially increase I_{OFF} , this because weak inversion state leakage and is the function of V_{th}.

- We explore all leakage mechanisms contributing to the off-state current as described in the below figure.

I_1 : Reverse-bias p-n junction leakage
I_2 : Sub-threshold leakage
I_3 : Oxide tunneling current
I_4 : Gate current due to hot-carrier injection
I_5 : GIDL leakage
I_6 : Channel punch-through



Sources of Leakage Current

p-n Junction Reverse-Bias Current

- Has two main components :

1. Minority carrier diffusion near the edge of the depletion region
2. Electron-hole pair generation

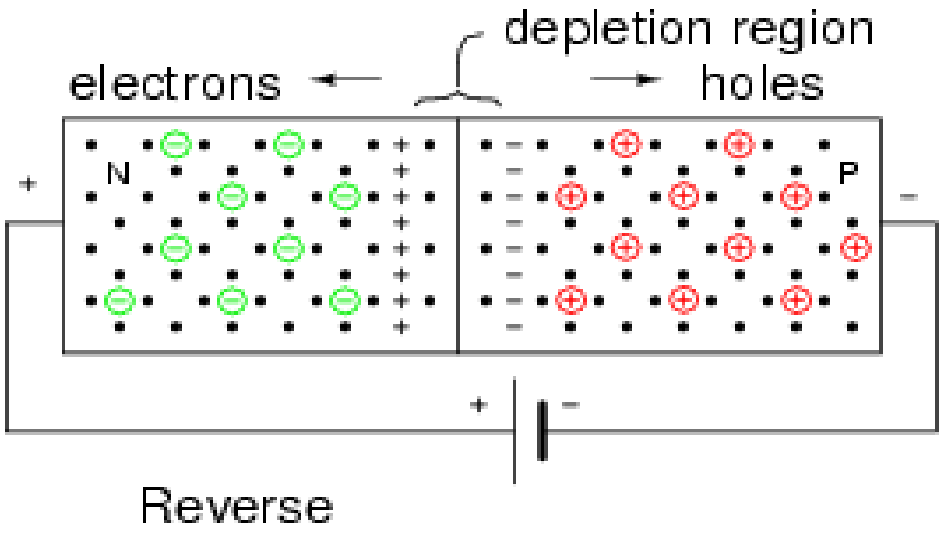


Fig. 1

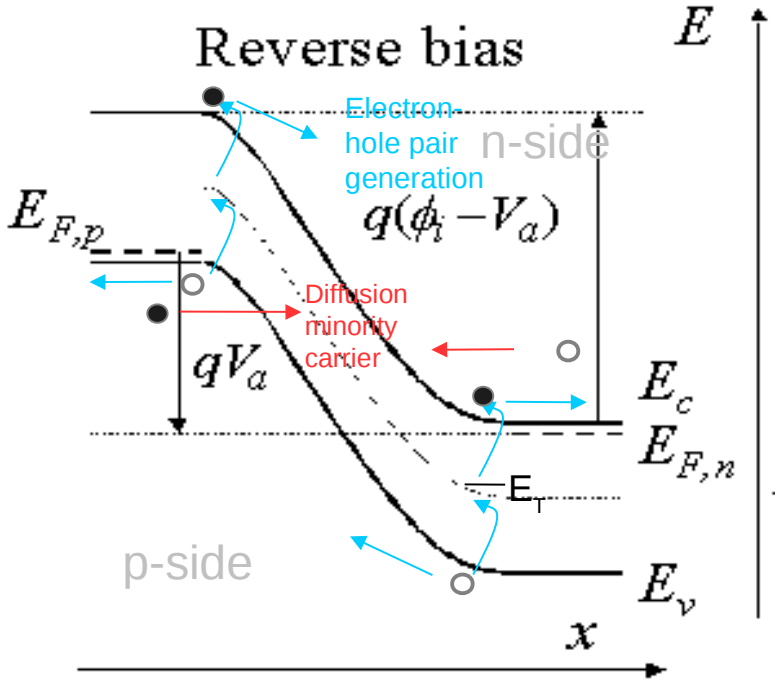


Fig. 2

pn Junction Reverse-Bias Current

- **If both n and p sides are heavily doped (for advaced MOSFETs: heavily doped shallow junctions and halo doping for better SCE(*))**

Band-to-Band tunneling(BTBT) dominates the pn junction leakage

This leakage mechanism is explain in next slide

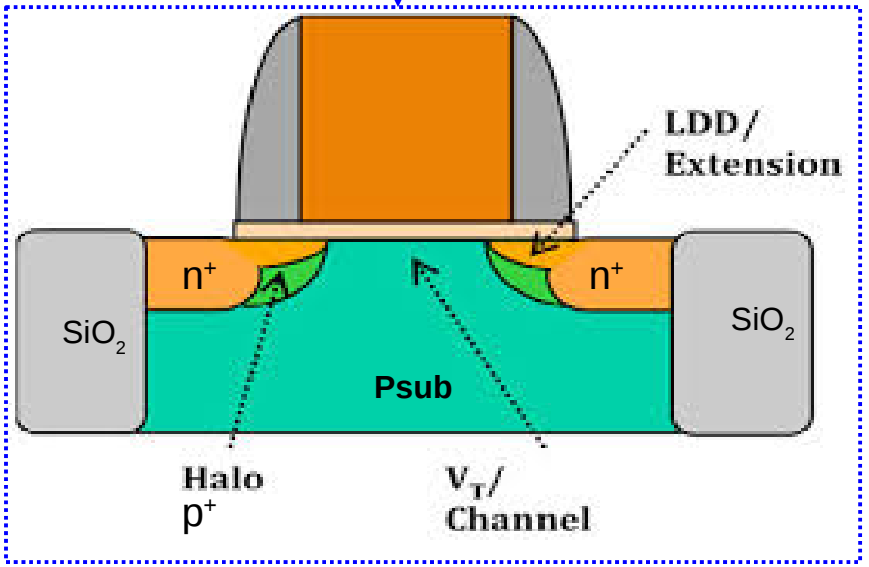


Fig. 3

(*) short channel effect

p-n Junction Reverse-Bias Current

Band-to-Band Tunneling Current

- It's evident that for the tunneling to occur(refer to fig.4) , total voltage across the junction has to be more than the band gap.

- The tunneling current density is expressed in the RHS formula.

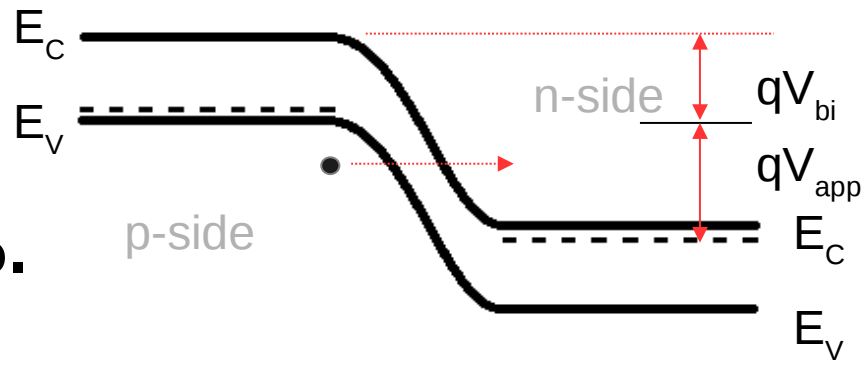


Fig. 4

$$J_{b-b} = A\frac{EV_{app}}{E_g^{1/2}}\exp\left(-B\frac{E_g^{3/2}}{E}\right)$$

$$A = \frac{\sqrt{2m^*}q^3}{4\pi^3\hbar^2}, \text{ and } B = \frac{4\sqrt{2m^*}}{3q\hbar}$$

$$E = \sqrt{\frac{2qN_aN_d(V_{app}+V_{bi})}{\varepsilon_{si}(N_a+N_d)}}$$

Where:

- m^* : effective mass of electron
- E_g : Band gap energy
- N_a, N_d : doping concentration in p and n sides
- V_{app} : applied voltage

(*) BTBT = band to band tunneling
RHS = Righ hand side
● = electron

Sub-threshold Leakage

General view

- Also called weak inversion conduction current (Vgs < V_{th}).

- In the weak inversion, the minority carrier concentration is small, but not zero.

- With both the number of mobile carriers and E(y) (fig.6) field are small -> drift current small.

- The sub-threshold conduction dominates by the diffusion current.

- Weak inversion typically dominates modern device off-state leakage due to the low V_{th}.

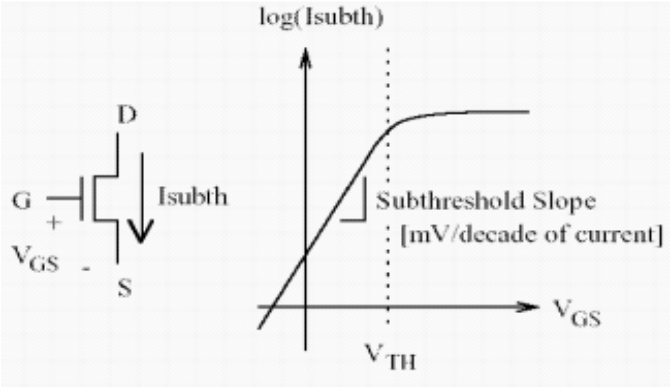


Fig. 5

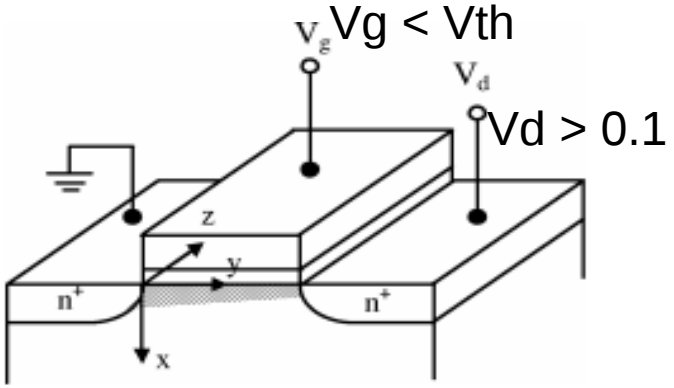


Fig. 6

Sub-threshold Leakage

General view

- The weak inversion current can be expressed:

$$I_{ds} = \mu_0 C_{ox} \frac{W}{L}(m-1)(v_T)^2 \times e^{(V_g - V_{th})/mv_T} \times \left(1 - e^{-v_{DS}/v_T}\right)$$

Where $$m = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{\frac{\varepsilon_{si}}{W_{dm}}}{\frac{\varepsilon_{ox}}{t_{ox}}} = 1 + \frac{3t_{ox}}{W_{dm}}$$

- We also consider the slope of $\log_{10}(I_{DS})$ vs. Vgs called substhreshold slope (S_t) given by.

$$S_t = \left(\frac{d(\log_{10} I_{ds})}{dV_{gs}}\right)^{-1} = 2.3\frac{mkT}{q}$$
$$= 2.3\frac{kT}{q}\left(1 + \frac{C_{dm}}{C_{ox}}\right).$$

- This show how effectively the transistor is be off when $V_{gs} < V_{th}$.

- From the equation, S_t can be made smaller by using thinner oxide layer or lower substrate doping concentration(lower S_t is desirable).

W_{dm} : depletion width under channel

Sub-threshold Leakage

Drain-induced Barrier Lowering (DIBL)

- In SC device, the source and drain depletion width and potential have a strong effect on the band bending over significant portion of the device => V_{th} and consequently the sub-threshold current varies with drain bias and this referred to as DIBL.

- DIBL occurs when depletion region of S, D interact with each other to lower S potential barrier.**

- DIBL is enhanced at high drain voltages and shorter channel length.

Curve	L	V_{ds}
A	6.25µm	0.5V
B	1.25µm	0.5V
C	1.25µm	5V

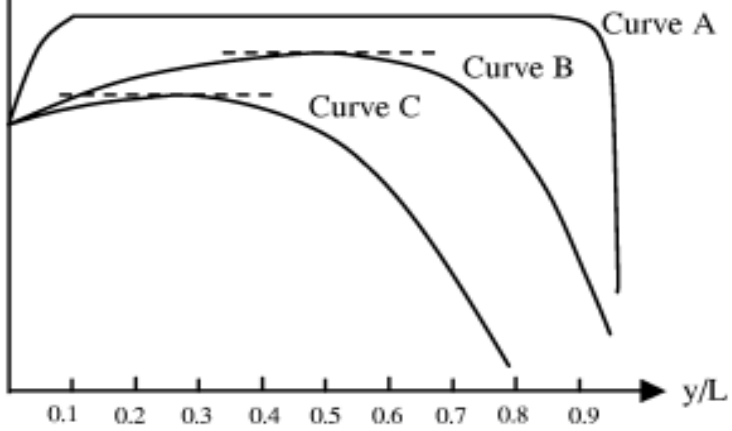


Fig. 7

** this causes source injects carriers into channel surface independent of gate voltage.

Sub-threshold Leakage
Drain-induced Barrier Lowering (DIBL)

-

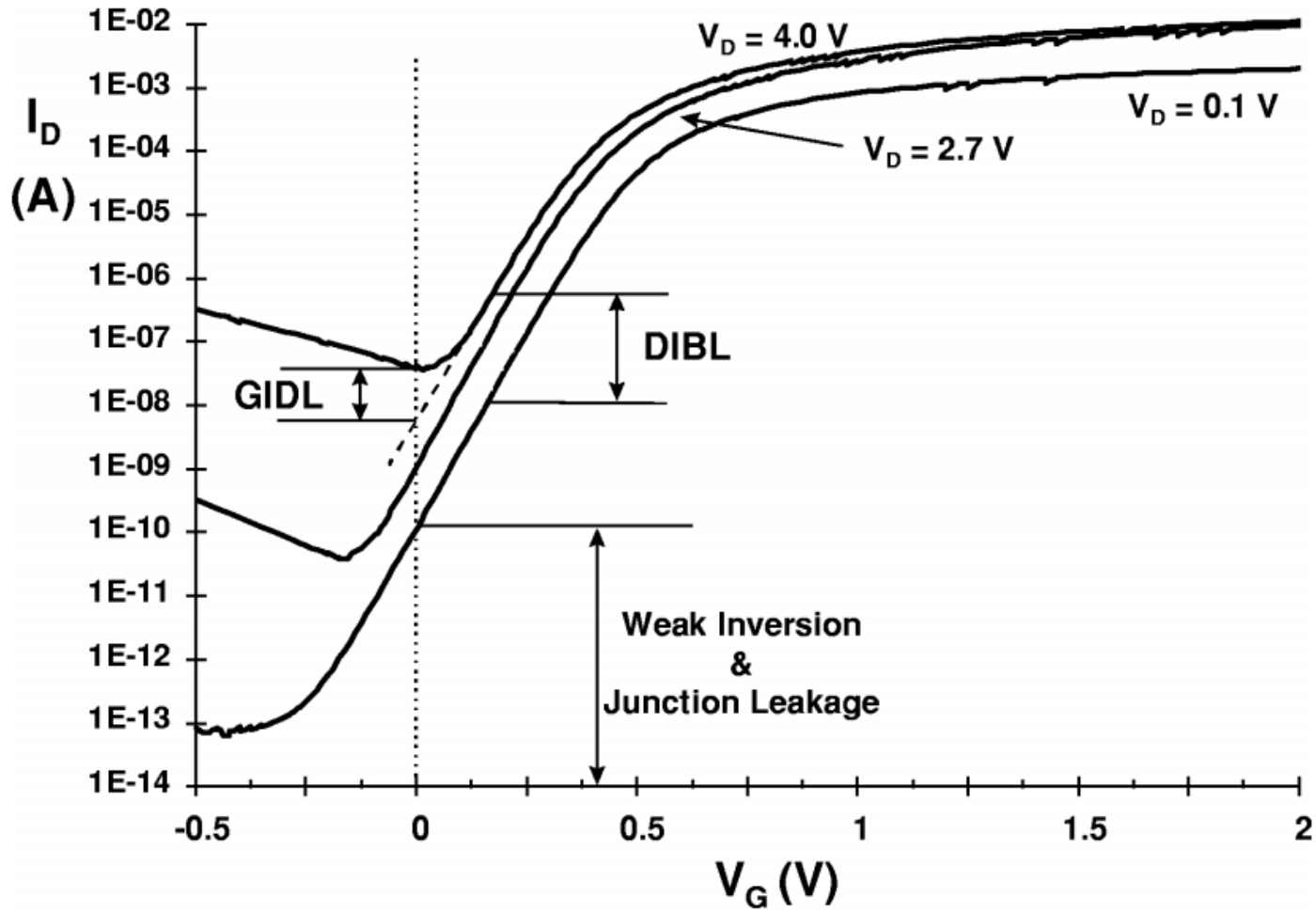


Fig. 7.1

N channel I_D vs. V_G showing DIBL, GIDL, weak inversion, and p-n junction reverse-bias leakage components

Sub-threshold Leakage
Body Effect

- **Reverse biasing well-source junction widens the bulk depletion region and increase V_{th}**

$$V_{\text{th}} = V_{\text{fb}} + 2\psi_B + \frac{\sqrt{2\varepsilon_{\text{si}} q N_a (2\psi_B + V_{\text{bs}})}}{C_{\text{ox}}} \quad \text{hence} \quad \frac{dV_{\text{th}}}{dV_{\text{bs}}} = \frac{\sqrt{\frac{\varepsilon_{\text{si}} q N_a}{2(2\psi_B + V_{\text{bs}})}}}{C_{\text{ox}}}$$

is substrate sensitivity

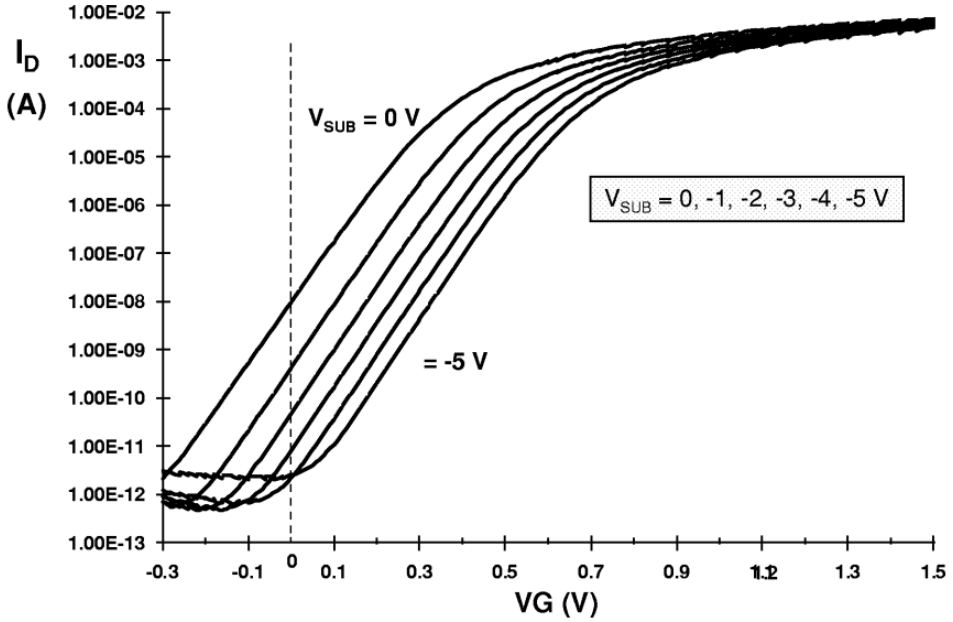
- On fig. 8, we can see that I_{OFF} decrease and V_{th} increase As Vsb increase

- Model of I_{sub} that include weak inversion, DIBL, body-effect :

$$I_{\text{subth}} = A \times e^{1/mv_T\left(V_G - V_S - V_{\text{th0}} - \gamma' \times V_S + \eta V_{\text{DS}}\right)} \times \left(1 - e^{-v_{\text{DS}}/v_T}\right)$$

where

$$A = \mu_0 C'_{\text{ox}} \frac{W}{L_{\text{eff}}} (v_T)^2 e^{1.8} e^{-\Delta V_{\text{th}}/\eta v_T}$$



N channel log(I_D) versus V_G for six substrate biases on a 0.35 um logic process technology (V_D = 2.7 V)

Fig. 8

Sub-threshold Leakage

Narrow-Width Effect

- Decreasing gate width modulates the threshold voltage of a transistor hence sub-threshold leakage.

- There are mainly three ways this effect change V_{th}. First we consider the fig.9.

- The existence of fringing field cause the gate-induced depletion to spread outside the defined channel width => increase total depletion charge.

- Moreover, we have: $V_{\text{th}} = V_{\text{fb}} + \phi_s + \frac{Q_B}{C_{\text{ox}}}$

Q_B increase by ΔQ_B so V_{th} increase, this amount can be modeled as:

$$V_{\text{NCE}} = \frac{\pi q N_{\text{sub}} x_{d,\max}^2}{2 C_{\text{ox}} W_{\text{eff}}} = 3\pi \frac{t_{\text{ox}}}{W_{\text{eff}}} \phi_s$$

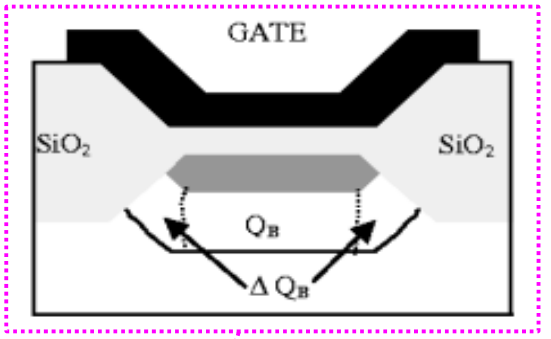
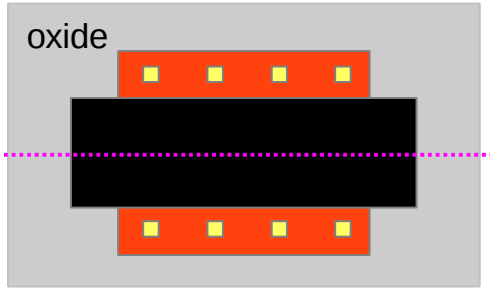


Fig. 9



Sub-threshold Leakage

Narrow-Width Effect

- **Second, Due to channel stop, dopants encroach under the gate. Hence, higher voltage is needed to completely invert the channel .**

- **A more complex effect is seen in trench isolation devices (fig.10) known as inverse-narrow-width effect.**

- **This structure eliminates increasing ΔQ_B => no increase of V_{th}**

- **On the other hand, due to edge-fringing effect at the gate edges, the inversion at the edges require less voltage than that in the center of gate.**

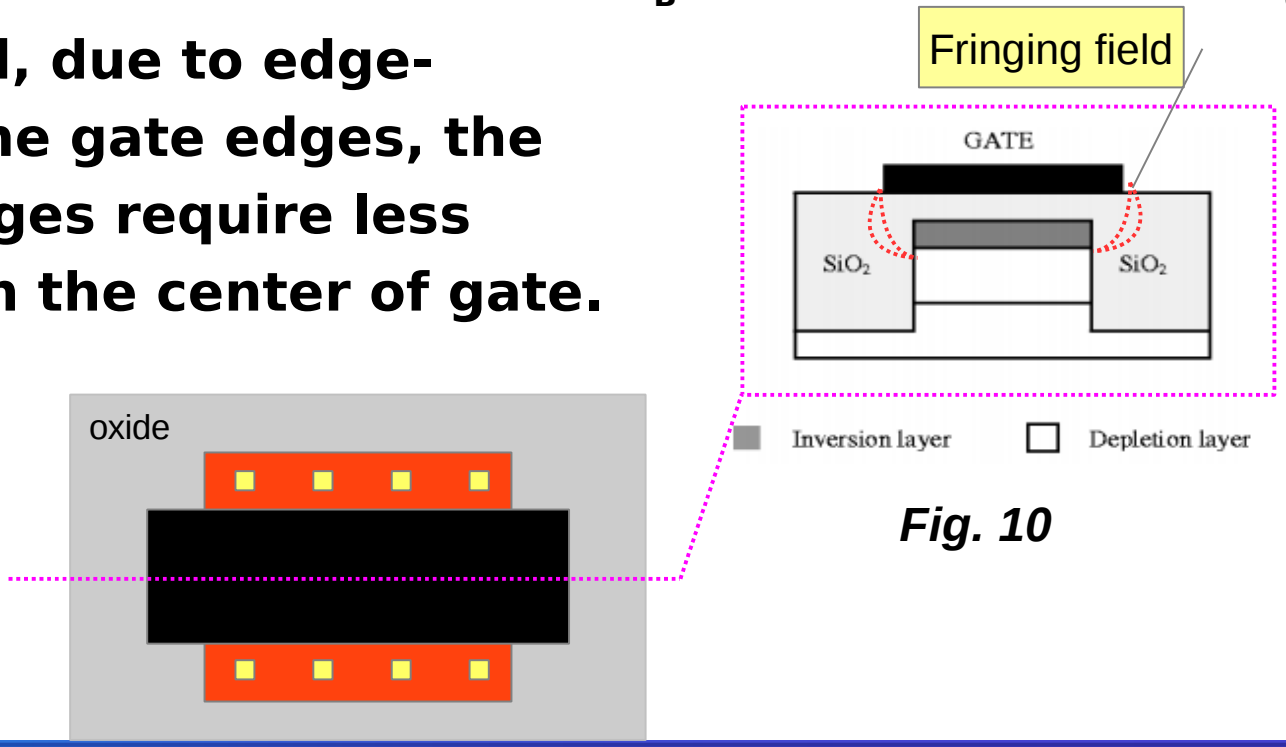


Fig. 10

Sub-threshold Leakage

Narrow-Width Effect

- **Moreover, gate capacitance now includes the side capacitance (C_F) due to overlap of the gate with the oxide isolation. $Cg = C_{ox}W + 2C_F$**
- **From equation :** $V_{\text{th}} = V_{\text{fb}} + \phi_s + \frac{Q_B}{C_{\text{ox}}}$
- **From above, the overall V_{th} reduce**

As shown in fig.12.

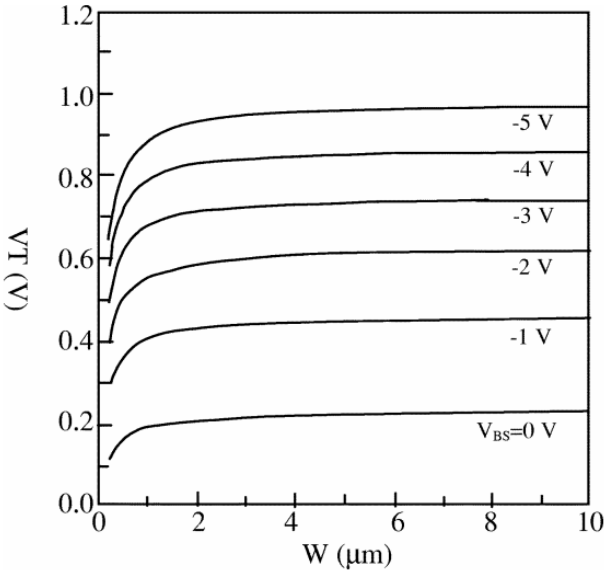


Fig. 12

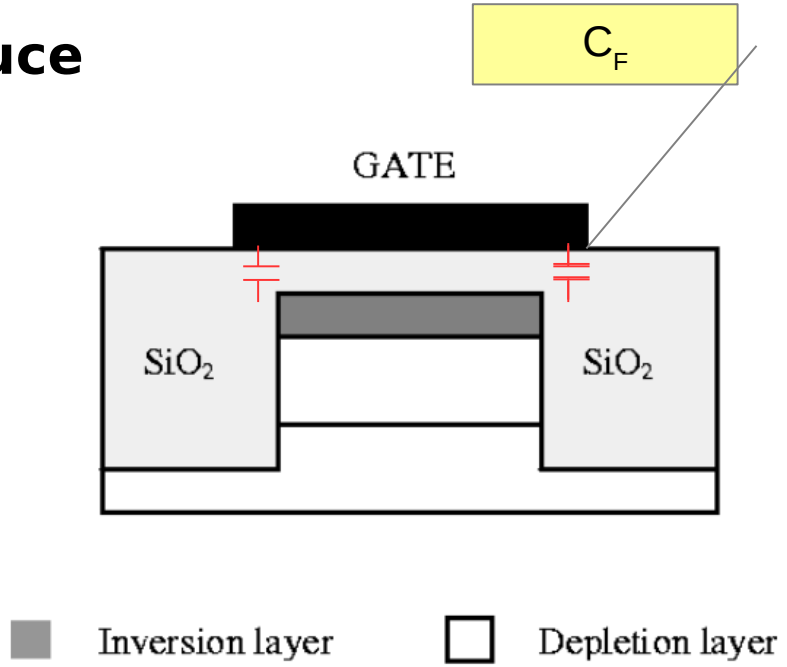


Fig. 11

Sub-threshold Leakage

Effect of Channel Length and V_{th} Rolloff

- Threshold voltage of MOSFET decreases as the channel length is reduced. This reduction of threshold voltage with reduction of channel length is known as V_{th} roll-off.

- In the case of short-channel devices, source-to-drain distance is comparable to the depletion width in the vertical direction.

- The source and drain depletion regions now penetrate more

into the channel length => need less voltage gate to invert bulk charge to turn on a transistor => V_{th} ↓

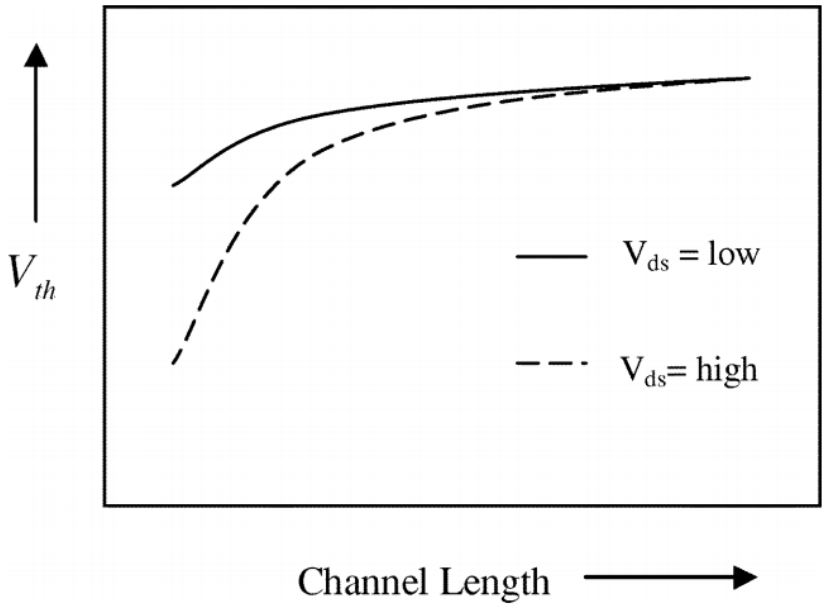
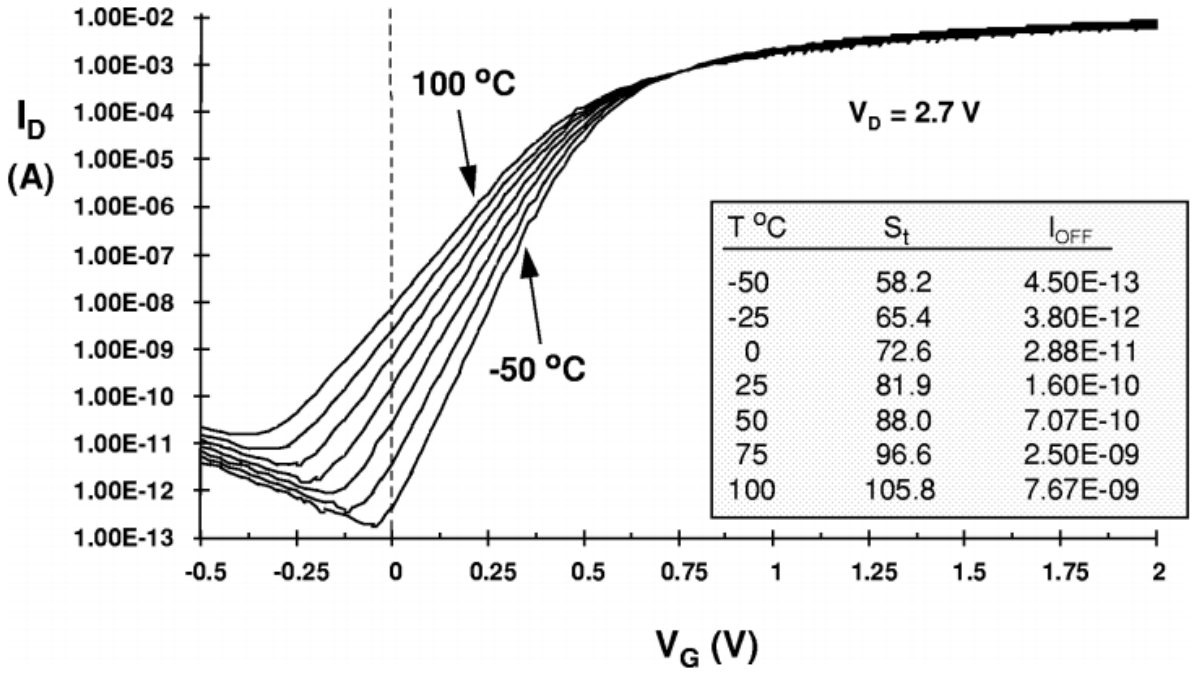


Fig. 13

Sub-threshold Leakage
Effect of Temperature

- Temperature dependence of the subthreshold leakage current is important, since VLSI circuits usually operate at elevated temperatures due to the power dissipation (heat generation) of the circuit.

- Fig. 14 show that S_t varies from 58.2 – 81.9 mV/dec as T^0 increases from -50 - 25 ^{0}C



T °C	S_t	I_{OFF}
-50	58.2	4.50E-13
-25	65.4	3.80E-12
0	72.6	2.88E-11
25	81.9	1.60E-10
50	88.0	7.07E-10
75	96.6	2.50E-09
100	105.8	7.67E-09

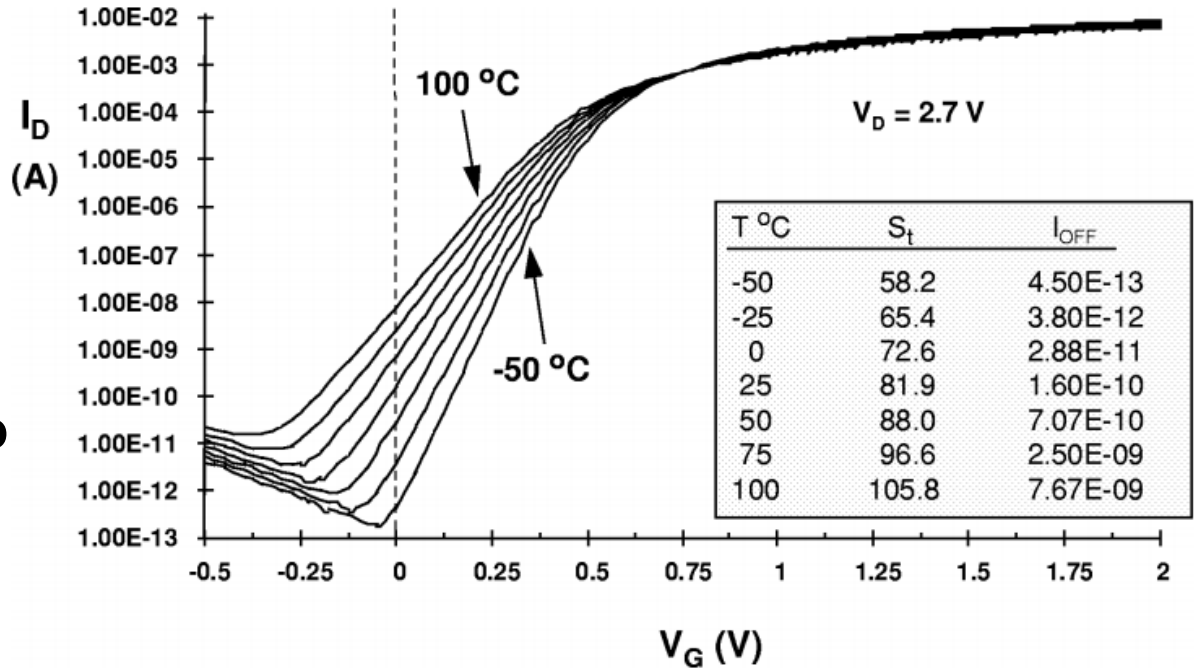
Fig. 14

Sub-threshold Leakage
Effect of Temperature

- Two parameters increase the sub-threshold leakage as temperature is raised.

1. **S_t linearly increases with T^0**
2. **V_{th} decreases**

- The temperature sensitivity of V_{th} about 0.8mV/^{0}C. This can be used to estimate I_{OFF} at other T^0 in this technolo



T °C	S_t	I_{OFF}
-50	58.2	4.50E-13
-25	65.4	3.80E-12
0	72.6	2.88E-11
25	81.9	1.60E-10
50	88.0	7.07E-10
75	96.6	2.50E-09
100	105.8	7.67E-09

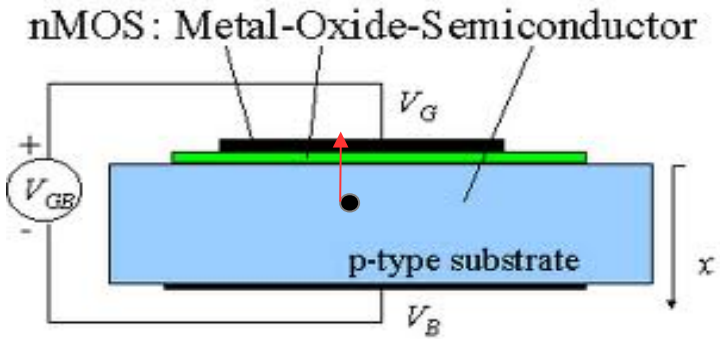
Fig. 14

Tunneling into and Through Gate Oxide

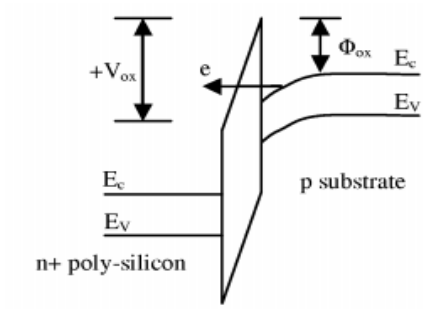
General View

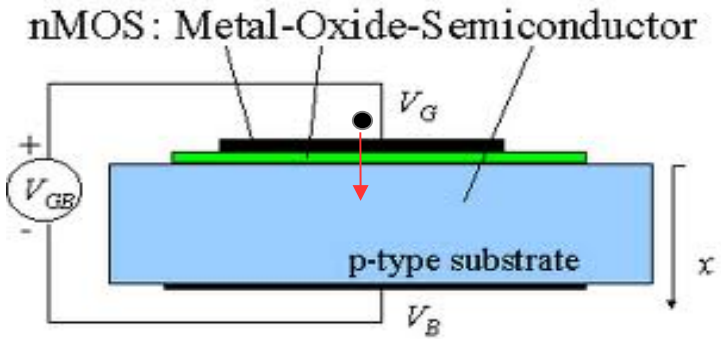
- Reduction of gate oxide thickness results in an increase in the field across the oxide results in tunneling of electrons from substrate to gate and vice verse, causing tunneling current.

For Example:



$V_{GB} > 0$





$V_{GB} < 0$

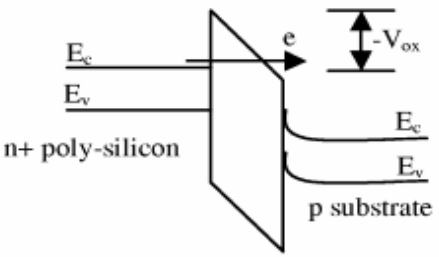
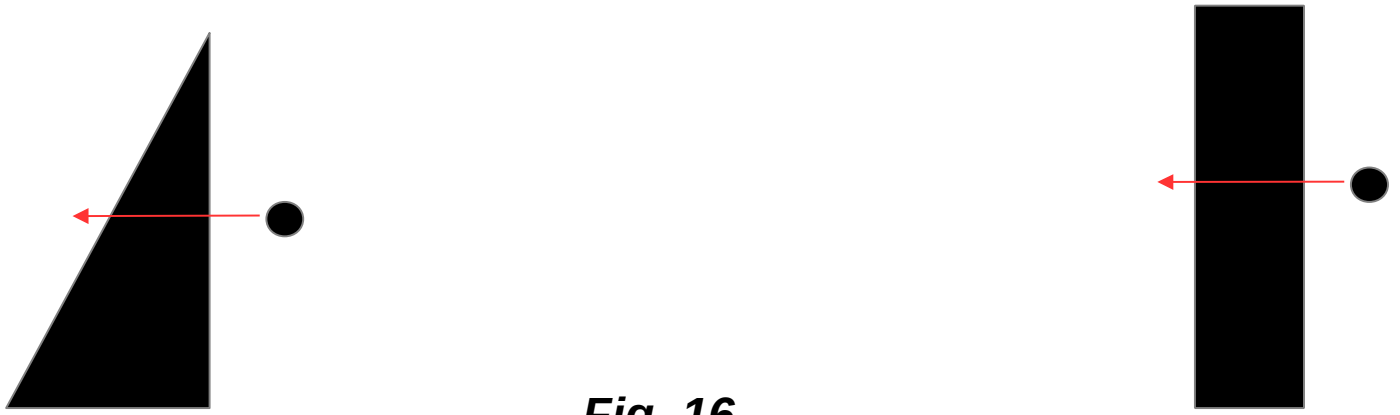


Fig. 15

Tunneling into and Through Gate Oxide

General View

- **The tunneling between substrate and gate can be divided into two parts, namely:**
  1. **Fowler-Nordheim (FN) tunneling**
  2. **Direct tunneling**
- **The tunneling probability of an electron depends on the thickness of the barrier, barrier height & the structure of barrier.**



(1) (2)

Fig. 16

Tunneling into and Through Gate Oxide

Fowler-Nordheim Tunneling

- In FN tunneling, electrons tunnel into the conduction band of the oxide layer

- Fig. 17 show FN tunneling of electrons from the inverted surface to the gate and represent as below:

$$J_{FN} = \frac{q^3 E_{ox}^2}{16\pi^2 \hbar \phi_{ox}} \exp\left(-\frac{4\sqrt{2m^*}\phi_{ox}^{3/2}}{3\hbar q E_{ox}}\right)$$

Where:

- $V_{ox} > \Phi_{ox}$ (valid for above equation)
- V_{ox} :Voltage drop across the oxide
- m^* : Effective mass of electron
- E_{ox} : Field across the oxide

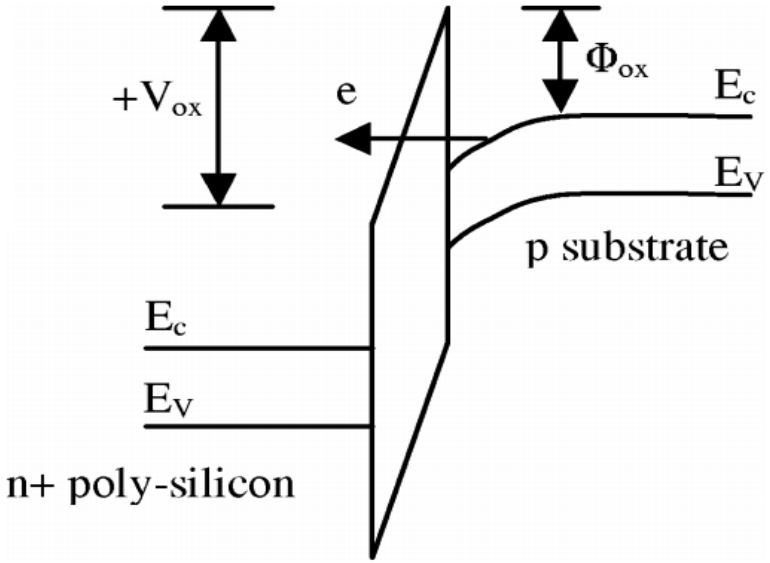


Fig. 17

Tunneling into and Through Gate Oxide

Fowler-Nordheim Tunneling

- The value of FN tunneling current quite small, for example: for E_{ox} = 8x10^{6} MV/cm then J_{FN} = 5x10^{-7} A/cm^{2}.

- Since Φ_{ox} = 3.1 eV, short-channel devices mostly operate at V_{ox} < Φ_{ox} => for normal device operation FN current is negligible.

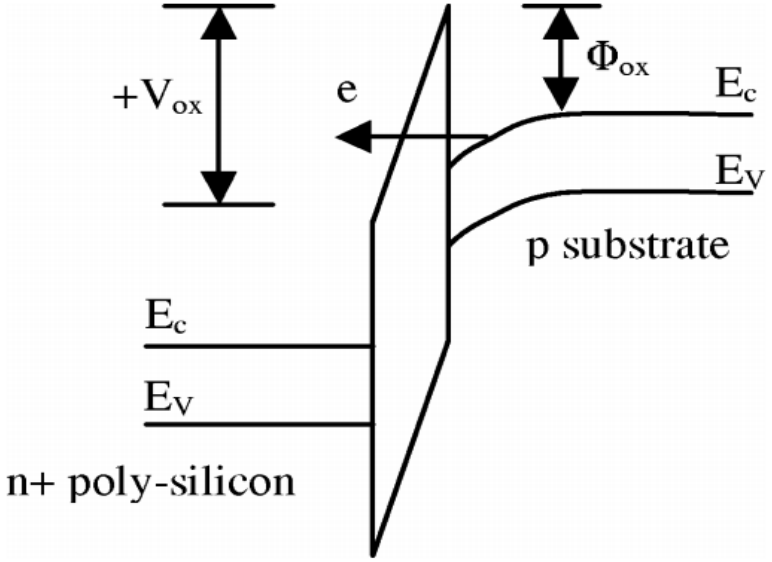


Fig. 17

Tunneling into and Through Gate Oxide

Direct Tunneling

- In very thin oxide layer (less than 3-4 nm), electron from inverted silicon surface directly tunnel to gate thru the forbidden energy gap of the SiO_2 layer.

- The direct tunneling occurs

At $V_{ox} < \Phi_{ox}$. Current density

Modeled as:

$$J_{DT} = AE_{ox}^2 \exp\left\{-\frac{B\left[1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right]}{E_{ox}}\right\}$$

where $A = q^3/16\pi^2\hbar\phi_{ox}$ and $B = 4\sqrt{2m^*}\phi_{ox}^{3/2}/3\hbar q$.

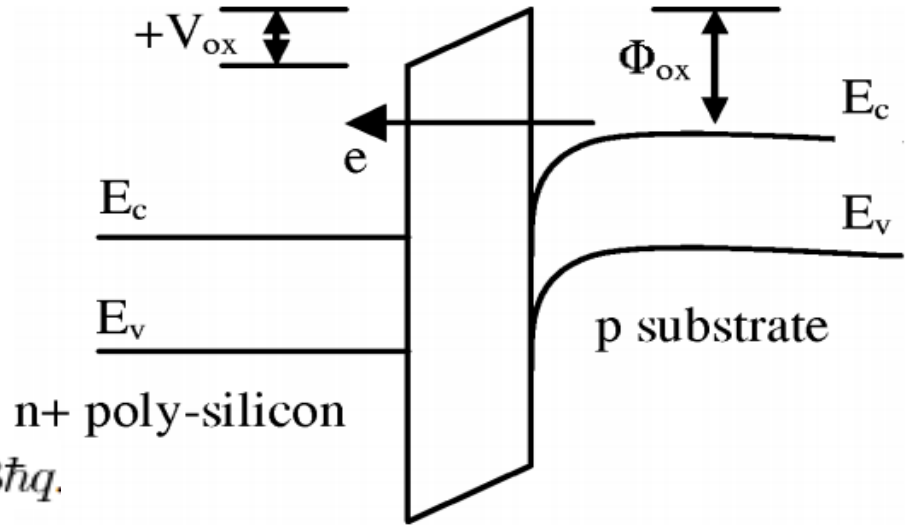


Fig. 18

Tunneling into and Through Gate Oxide

Direct Tunneling : Components of tunneling current

- The gate direct tunneling current can be divided into five major components as described in fig. 18.

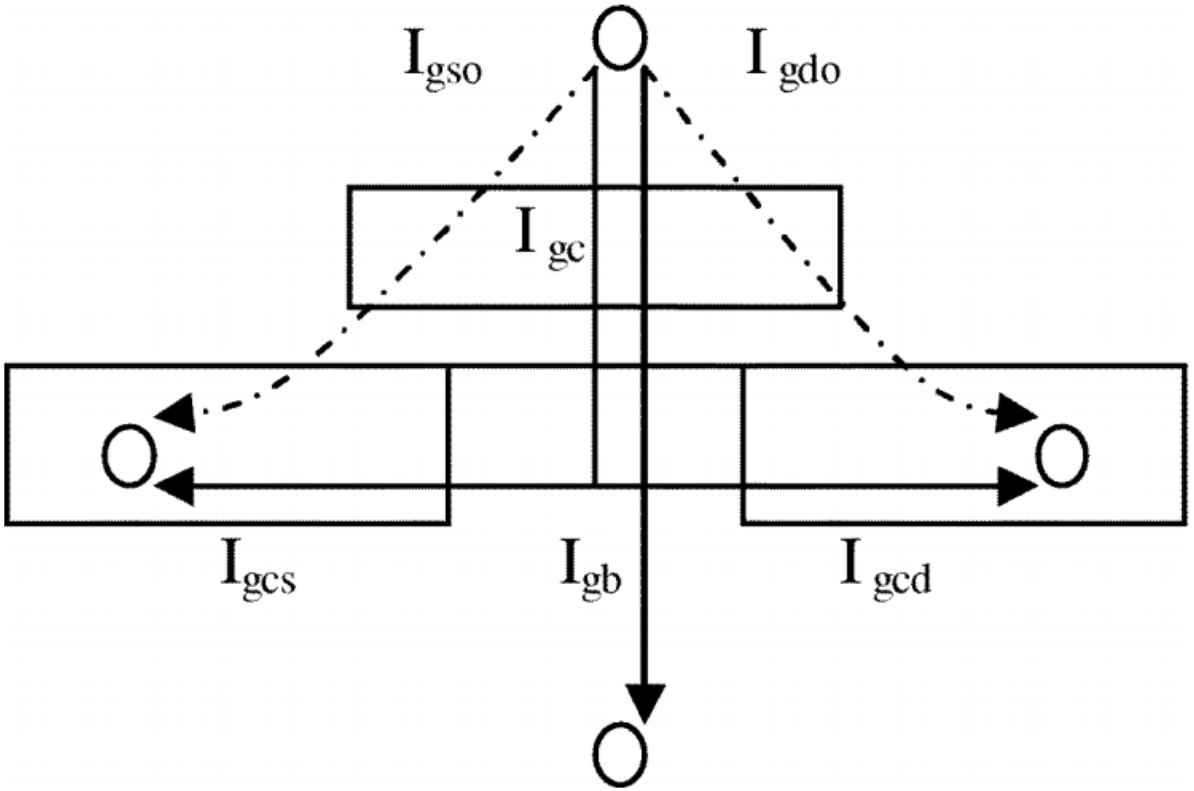


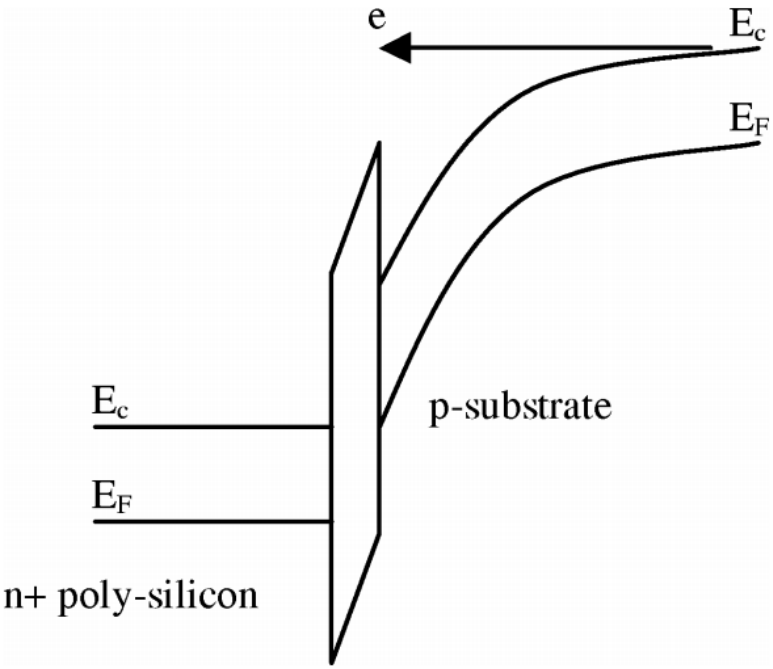
Fig. 18

Injection of Hot Carriers from Substrate to Gate Oxide

General View

- In a short-channel transistor, due to high electric field near the Si–SiO_2 interface, electrons or holes can gain sufficient energy from the electric field to cross the interface potential barrier and enter into the oxide layer as described in fig.19

- The injection is more likely for electrons than hole since electrons has lower effective mass than Holes and barrier height for holes (4.5 eV), (3.1 eV) for electrons.



Injection of hot electron from substrate to oxide

Fig. 19

Gate-Induced Drain Leakage(GIDL)

General View

- When the gate is biased to form an accumulation layer at silicon surface, the presence of hole cause the surface behaves like p^{++} region => the depletion of this layer is smallest.

- When Vg large (below zero)

the n+ drain region under

the gate can be depleted

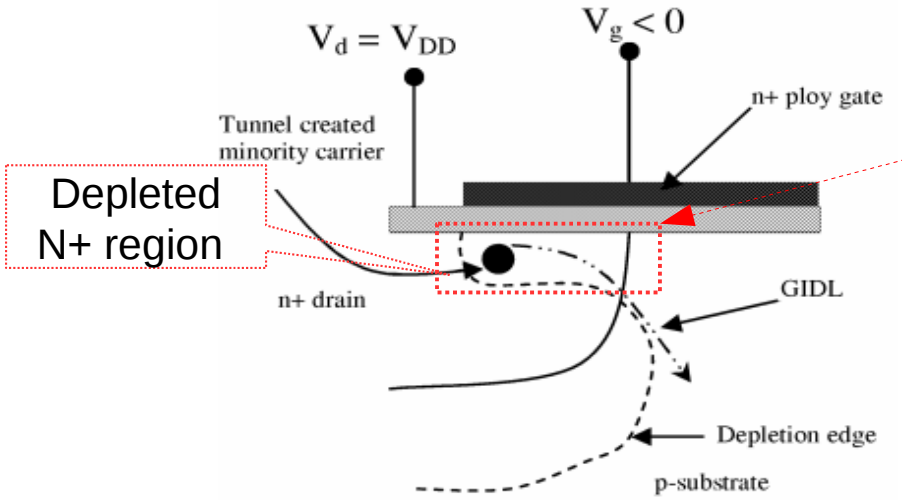


Fig. 21

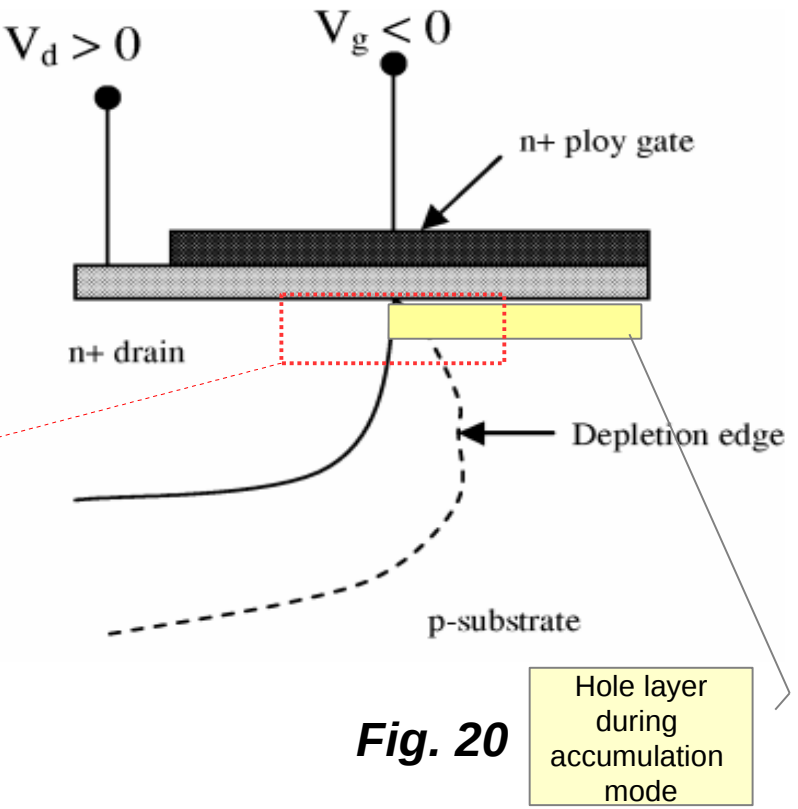


Fig. 20

Gate-Induced Drain Leakage
General View

- As a result of all these effects, minority carriers(tunneling ones, hot carriers) are emitted in the drain n+ region underneath the gate.

- Since the substrate is at a lower potential for minority carriers, these carriers underneath the gate are swept laterally to substrate.

- Thinner oxide thickness,

Higher VDD => increase GIDL

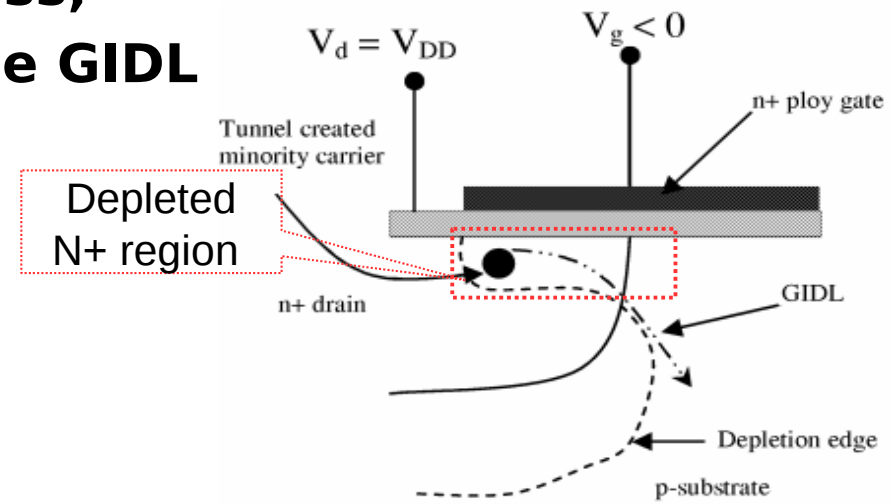


Fig. 21

Punch-through
General View

- When the combination of shorter channel length and reverse bias leads to the merging of the depletion regions, punch-through is said to have occurred as show in fig. 22.

- A V_{th} adjust implant is used to have a higher doping at the surface than bulk's => greater expansion of depletion layer below the surface. Thus, the punch-through occurs below the surface.

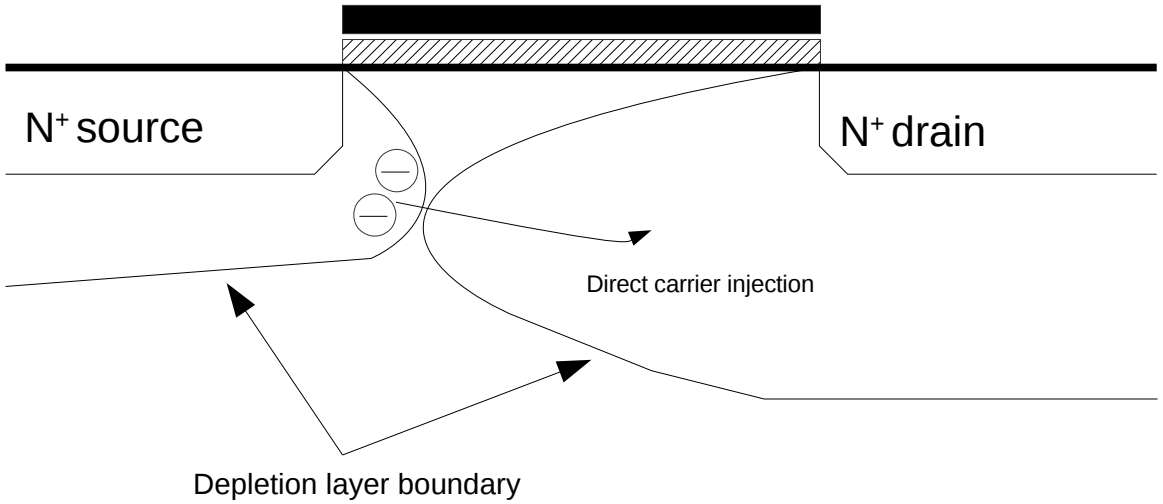


Fig. 22

Punch-through

General View

- The punch-through lower the potential barrier for the majority carriers in the source => more of these carriers cross the barrier and enter substrate, then the drain collects some of them.

- V_{PT} estimates the value of V_{DS} for which the punch-through occurs at V_{gs} = 0.

$$V_{\mathrm{PT}} \propto N_B(L - W_j)^3$$

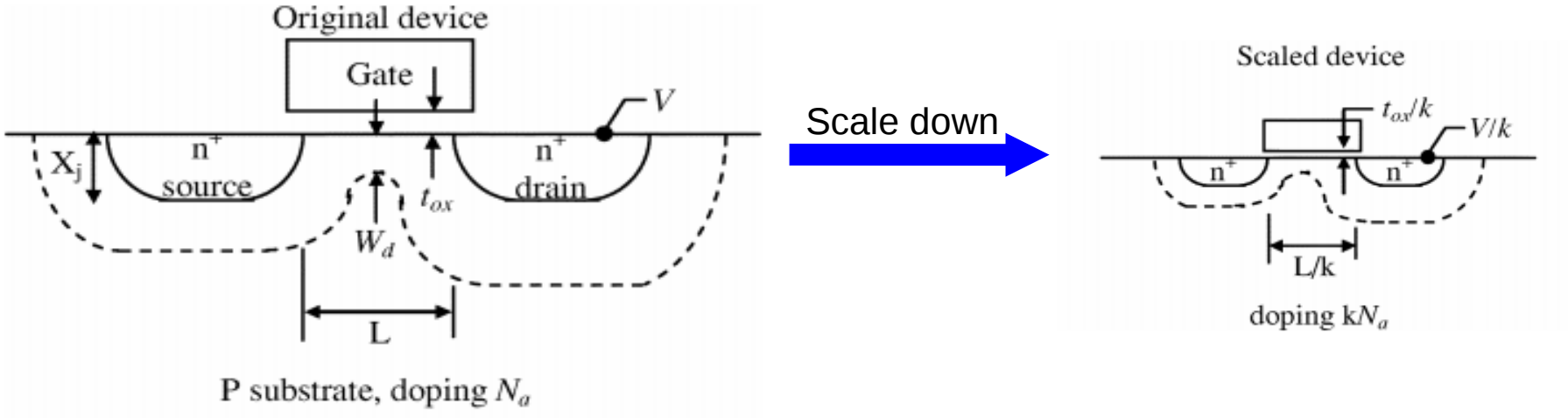
Where:

- N_B : bulk doping concentration
- L : channel length
- W_J : junction width

Leakage Reduction Technique

Channel Engineering for Leakage Reduction
General View

- Based on constant field scaling, the SCE can be kept under control by scaling down the vertical dimensions, e.g. gate oxide thickness, junction depth, along with horizontal dimensions, also reduce supply voltage, increase bulk doping concentration accordingly described in blow figure.



MOSFET constant electric field scaling
Fig. 23

Channel Engineering for Leakage Reduction

General View

- The goal is to optimize the channel profile to minimize the OFF-state leakage while maximizing the linear and saturated drive currents.

- Super-steep retrograde wells and halo implants have been used as a means to scale the channel length and increase the transistor drive current without causing an increase in the OFF-state leakage current

- Fig. 24 is a schematic representation of the transistor regions that are affected by the different types of well engineering

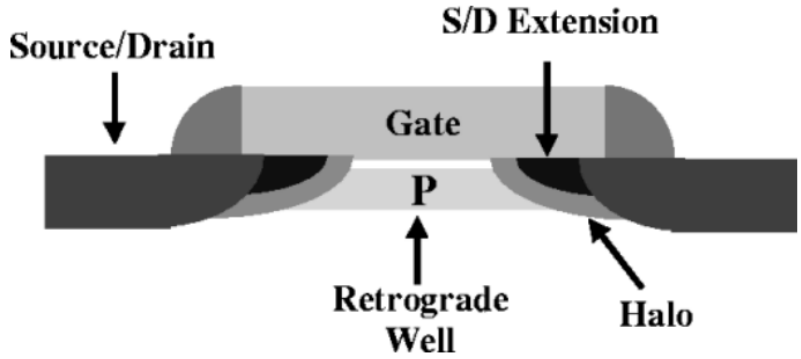


Fig. 24

Channel Engineering for Leakage Reduction

Retrograde Doping

- To maintain acceptable OFF-state leakage with continually decreasing channel lengths, both the oxide thickness and the gate-controlled depletion width in silicon must be reduced in proportion to the channel length(L) to offset the degradation in SCEs for extremely small devices.

$$W_{\mathrm{dm}} = \sqrt{\frac{4\varepsilon_{\mathrm{si}}\psi_B}{qN_a}}$$

- This require increase Na (causing V_{th} ↑)

$$V_{\mathrm{th}} = V_{\mathrm{fb}} + 2\psi_B + \frac{\sqrt{4\varepsilon_{\mathrm{si}}qN_a\psi_B}}{C_{\mathrm{ox}}}$$

(*) W_{dm} : depletion width

SCE : short channel effect

Channel Engineering for Leakage Reduction
Retrograde Doping

- If V_{th} is not scaled down, the device performance in lower supply voltage will degrade => to achieve both reducing W_{dm} and V_{th} retrograde doping can be used.

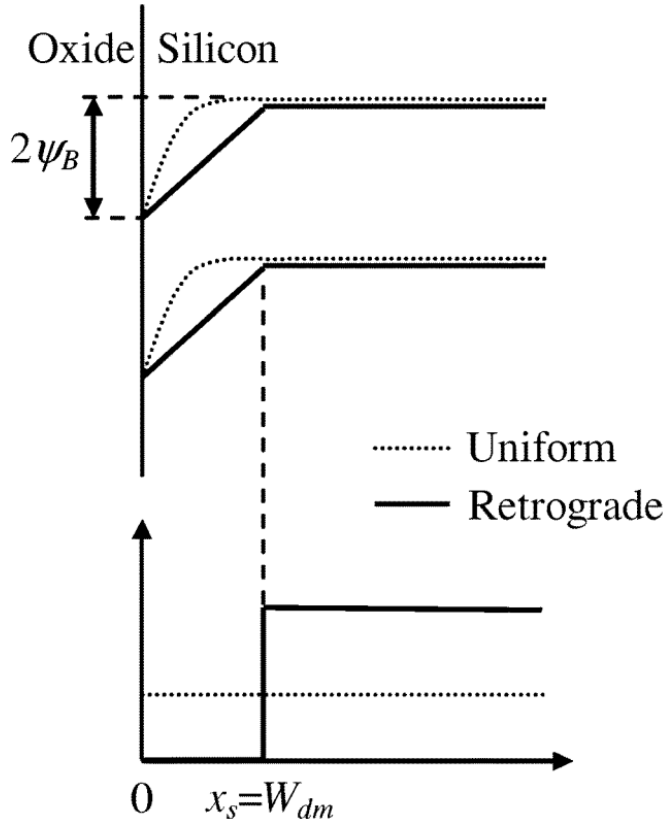
- This is used to improve SCEs and to increase surface mobility by creating a low surface channel concentration, highly doped subsurface region act as a barrier against punch-through.

(*) W_{dm} : depletion width

Channel Engineering for Leakage Reduction

Retrograde Doping

- As for fig. 25, we see that, the depletion charge and electric field of extreme retro-grade is 1/2 that of a uniform doped channel => reduce V_{th} and improve mobility.



Band diagrams (shown on top) at the threshold condition for a uniformly doped and an extreme retrograde-doped channel (doping profiles shown at bottom)

Fig. 25

(*) W_{dm} : depletion width

Channel Engineering for Leakage Reduction

Halo Doping

- Halo doping or nonuniform channel profile in a lateral direction was introduced below 0.25 um technology node to provide another way to control the dependence of threshold on channel length.

- Halo doping areas is created by injecting point defects during sidewall oxidation, these defect collect doping impurities from bulk => increase doping concentration near source, drain end of the channel.

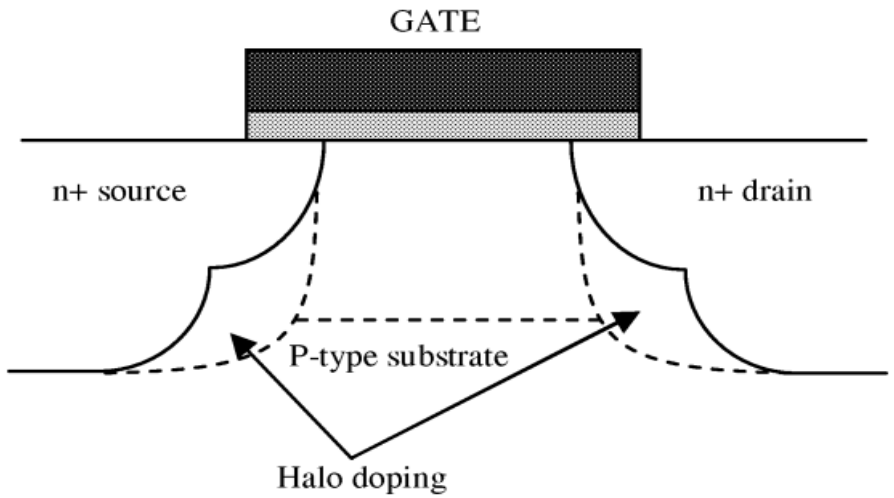


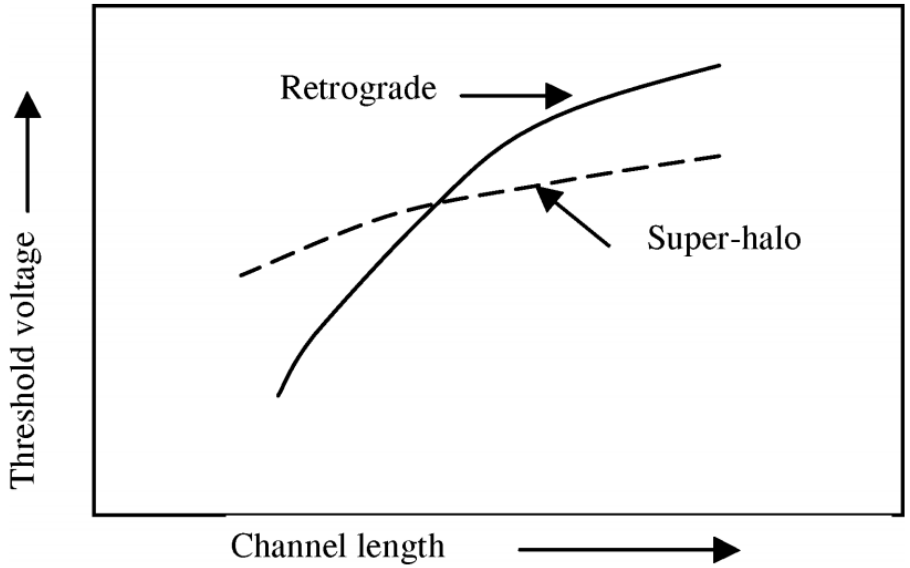
Fig. 26

Channel Engineering for Leakage Reduction

Halo Doping

- More highly doped p-type substrate near the edges of the channel reduces the charge-sharing effects from the source and drain fields => reducing the width of the depletion region in the drain-bulk and source-bulk regions.

- Reduction of charge-sharing effects reduces the threshold voltage degradation due to channel length reduction as shown in fig. 27.



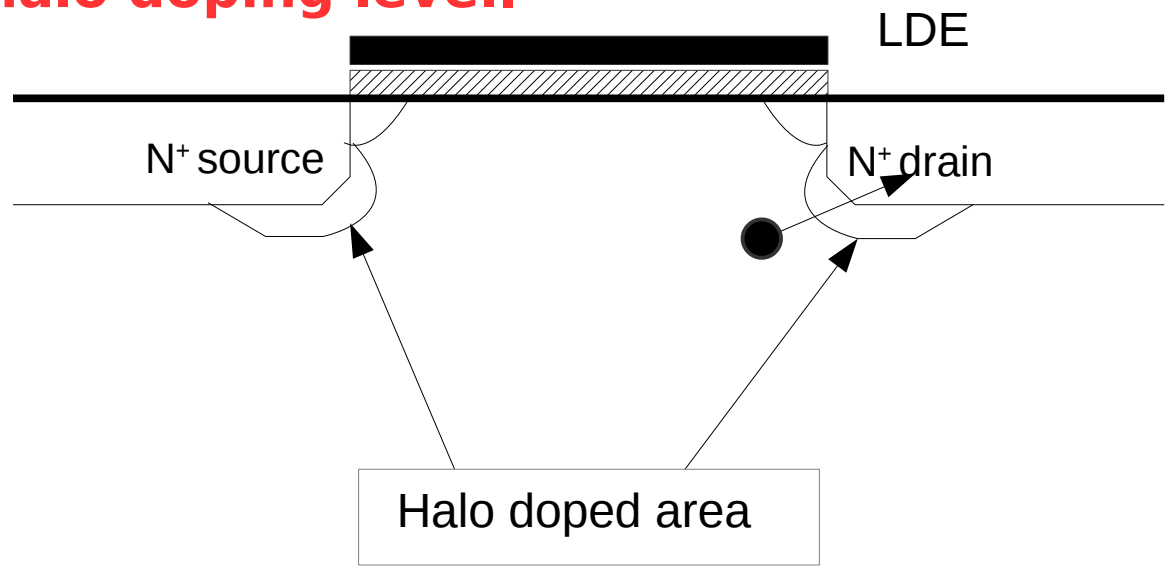
Short-channel threshold-voltage roll-off for retrograde and super-halo (vertical and lateral nonuniform doping).

Fig. 27

Channel Engineering for Leakage Reduction
Halo Doping

- The reduction in drain and source junction depletion region width also reduces the barrier lowering in the channel, thus reducing DIBL, the punch-through possibility.

- Also, The BTBT currents in the high-field region near the drain ultimately limit the halo doping level.



Conclusions

Conclusions

- In current deep-submicron devices with low threshold voltages, sub-threshold and gate leakage have become dominant sources of leakage and are expected to increase with the technology scaling.
- GIDL and BTBT may also become a concern in advanced CMOS devices
- To manage the increasing leakage in deep-submicron CMOS circuits, solutions for leakage reduction have to be sought both at the process technology and circuit levels.
- At the process technology level, well-engineering techniques by retrograde and halo doping are used to reduce leakage and improve short-channel characteristics.

Conclusions

- There many techniques to reduce leakage current but not mentioned here, those are (at the circuit level consideration)
  - Standby Leakage Control Using Transistor Stacks (Self-Reverse Bias)
  - Multi V_{th} design
  - Dynamic V_{th} design

Thank you for your intention!

©2014 Renesas Electronics Corporation Circuit Design Group