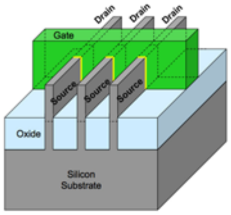
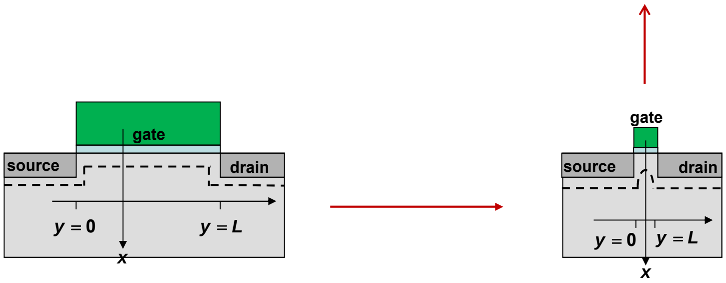
Lecture 26

Nano-Scale FETs

In this lecture you will learn:

- Nano-scale FETs
- FET size scaling
- Short channel length effects
- Fin-FETs





Why Nano-Scale FETs?

- Smaller FETs (shorter channel lengths L) are faster (in both analog and digital applications):

  Smaller FETs will have a larger current I_D, larger g_m, smaller capacitances C_{gs} and C_{gd} , higher f_T, shorter rise and fall times t_r and t_f , all for the same voltages V_{GS} and V_{DD}

- Smaller FETs occupy smaller areas

  This means more FETs can be put in a chip

- Since power dissipation in a digital chip goes as V_{DD}^2, the higher performance of smaller FETs can be traded-off for lower power dissipation by scaling down the voltages

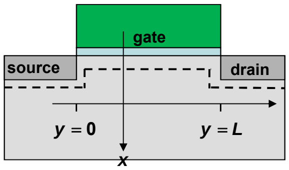
- New physics emerging at nano-scales enables one to design better FETs (…..but can also create new problems)

  Quantum transport, wave mechanics, energy level quantization, mobility engineering, ballistic scattering-free transport, coulomb blockade, …….

How does one scale down the size of a FET?

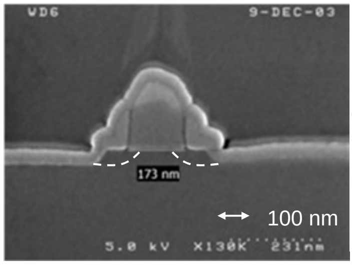
The Poisson Equation and the FET

The gradual channel approximation: Electrostatics in the vertical (x) direction are decoupled from the electrostatics in the horizontal (y) direction

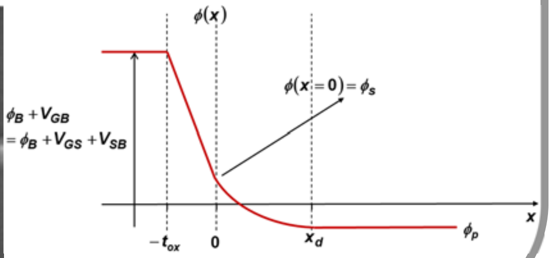


The “gradual channel approximation” works for a large channel length MOS transistors:

$$\frac{\partial^2\phi(x)}{\partial x^2}=-\frac{\rho}{\varepsilon}$$

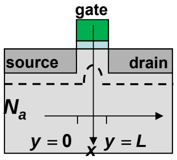


TEM of a 100 nm gate length MOS transistor



The Poisson Equation and the FET

Gradual channel approximation fails for a sub-micron MOS transistor – need to solve a 2-dimensional Poisson equation

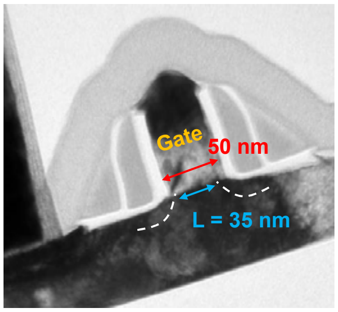


$$\frac{\partial^2\phi(x,y)}{\partial x^2}+\frac{\partial^2\phi(x,y)}{\partial y^2}=-\frac{\rho(x,y)}{\varepsilon}$$

The charge density is approximately that associated with the depletion regions (at least below threshold):

$$\rho(x,y)=-qN_a(x,y)$$

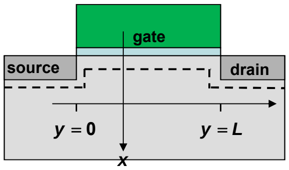
What new physics does a 2D Poisson equation produces?



TEM of a 35 nm gate length FET

Scaling Down the FET

$$\frac{\partial^2\phi(x,y)}{\partial x^2}+\frac{\partial^2\phi(x,y)}{\partial y^2}=\frac{qN_a(x,y)}{\varepsilon}$$

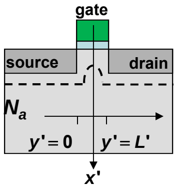


Suppose we:

1) Scale down all the dimensions by "λ"

$$x'=\frac{x}{\lambda}\qquad y'=\frac{y}{\lambda}$$

$$\Rightarrow\frac{\partial^2\phi(x',y')}{\partial x'^2}+\frac{\partial^2\phi(x',y')}{\partial y'^2}=\frac{qN_a(x',y')}{\varepsilon}\lambda^2$$



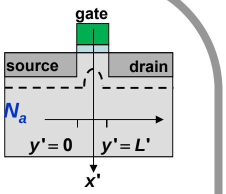
2) Scale down all the potentials by "κ"

$$\phi'(x',y')=\frac{\phi(x',y')}{\kappa}$$

$$\Rightarrow\frac{\partial^2\phi'(x',y')}{\partial x'^2}+\frac{\partial^2\phi'(x',y')}{\partial y'^2}=\frac{qN_a(x',y')}{\varepsilon}\left(\frac{\lambda^2}{\kappa}\right)$$

Scaling Down the FET

$$\Rightarrow\frac{\partial^2\phi'(x',y')}{\partial x'^2}+\frac{\partial^2\phi'(x',y')}{\partial y'^2}=\frac{qN_a(x',y')}{\varepsilon}\left(\frac{\lambda^2}{\kappa}\right)$$

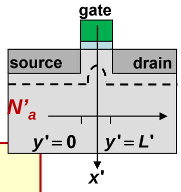


And suppose we:

3) Scale up all the dopings by "θ"

$$N_a'(x',y')=\theta\,N_a(x',y')$$

$$\Rightarrow\frac{\partial^2\phi'(x',y')}{\partial x'^2}+\frac{\partial^2\phi'(x',y')}{\partial y'^2}=\frac{qN_a'(x',y')}{\varepsilon}\left(\frac{\lambda^2}{\theta\,\kappa}\right)$$



If we choose: $\frac{\lambda^2}{\theta\,\kappa}=1$

then the 2D spatial potential profiles in the scaled device (whose dimensions are smaller by λ) would be the same as that in the prescaled device (however, the potentials would be scaled down by κ)

The electric field in the scaled device would be larger by a factor: $\frac{\lambda}{\kappa}$

$$\frac{\partial\phi'}{\partial x'}=\frac{\lambda}{\kappa}\frac{\partial\phi}{\partial x}\qquad\frac{\partial\phi'}{\partial y'}=\frac{\lambda}{\kappa}\frac{\partial\phi}{\partial y}$$

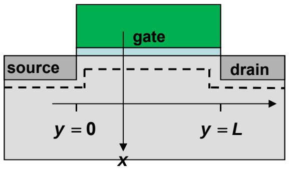
Scaling Down the FET: Problems

Constant electric field scaling:

If we choose:

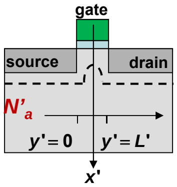
$$\lambda = \kappa = \theta$$

then the electric field in the scaled device would be the same as in the pre-scaled device



End result:

1) We reduced the device dimensions (good for faster operation)
2) We reduced the potentials (good for lower power dissipation)
3) We kept the device electric field the same (so no device breakdown problems in the smaller device)

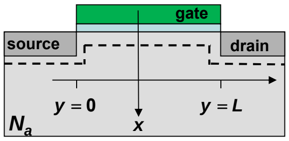


Problems with the device scaling described above:

1) Potentials cannot be scaled down too much when the device dimensions are reduced (need to leave enough room for noise margins, for example)
2) Consequently, electric field increases in smaller devices (and causes problems)
3) Material constants, like the potential ϕ_M of the gate metal and built-in potentials ϕ_B, and thermal voltage *KT/q* do not scale when the external applied voltages are scaled
4) Increase in device doping in smaller devices causes carrier mobility degradation

=> Smaller devices suffer from many problems (termed "short channel effects")

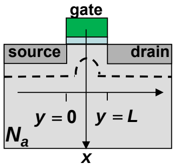
The Nano-Scale FET: 2D Poisson Equation



In the gradual channel approximation we had obtained:

$$\frac{\partial^2 \phi(x,y)}{\partial x^2} = \frac{qN_a}{\varepsilon_s} \quad \Big\{ \text{In the bulk}$$

$$V_{TN} = V_{FB} - 2\phi_p + \frac{\sqrt{2\,\varepsilon_s q N_a \left(-2\phi_p + V_{SB}\right)}}{C_{ox}}$$



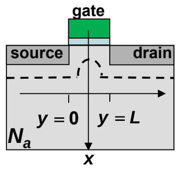
In nano-scale FET we need to consider the 2D Poisson equation:

$$\frac{\partial^2 \phi(x,y)}{\partial x^2} + \frac{\partial^2 \phi(x,y)}{\partial y^2} = \frac{qN_a}{\varepsilon_s}$$

$$\Rightarrow \frac{\partial^2 \phi(x,y)}{\partial x^2} = \frac{qN_a}{\varepsilon_s} - \frac{\partial^2 \phi(x,y)}{\partial y^2}$$

$$\Rightarrow \frac{\partial^2 \phi(x,y)}{\partial x^2} = \frac{q\left(N_a - \dfrac{\varepsilon_s}{q}\dfrac{\partial^2 \phi(x,y)}{\partial y^2}\right)}{\varepsilon_s} = \frac{qN_{a-eff}}{\varepsilon_s}$$

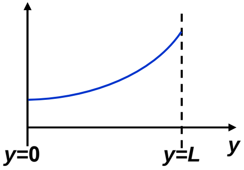
Threshold Voltage Roll-Off (A Short Channel Effect)



$$\frac{\partial^2\phi(x,y)}{\partial x^2}=\frac{qN_{a-eff}}{\varepsilon_s}$$

$$N_{a-eff}=N_a-\frac{\varepsilon_s}{q}\underbrace{\frac{\partial^2\phi(x,y)}{\partial y^2}}_{\text{Curvature}}<N_a$$

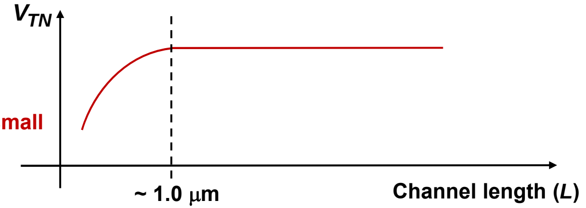
$\phi_s(y) = -\phi_p + V_{CS}(y)$



As the channel length is reduced, the potential curvature in the y-direction increases, and the effective doping is reduced

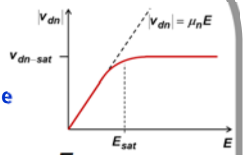
Since: $$V_{TN}=V_{FB}-2\phi_p+\frac{\sqrt{2\,\varepsilon_s qN_{a-eff}\left(-2\phi_p+V_{SB}\right)}}{C_{ox}}$$

The threshold voltage of the device reduces as the channel length is reduced!



Difficult to turn small FETs off!

Current Saturation Via Velocity Saturation (A Short Channel Effect)



Current saturation in long channel length (> 1 μm) FETs is due to pinch-off

In short channel length (< 1 μm) FETs, something else saturates the current before pinch-off; velocity saturation!

Recall that the drift velocity for electrons (handout 9) is: $$v_{dn}=-\frac{\mu_n E}{1+\frac{|E|}{E_{sat}}}$$

In Silicon: $E_{sat}\sim 5\times10^4$ V/cm

And the FET current is (handout 9):

$$I_D=W\,Q_N(y)\frac{\mu_n\,E(y)}{1+\frac{|E|}{E_{sat}}}=W\,C_{ox}\left(V_{GS}-V_{TN}-V_{CS}\right)\frac{\mu_n\frac{dV_{CS}(y)}{dy}}{1+\frac{1}{E_{sat}}\left|\frac{dV_{CS}(y)}{dy}\right|}$$

In the linear region, the above integrates to (handout 9):

$$I_D=\frac{W}{L}\mu_n\,C_{ox}\left(V_{GS}-V_{TN}-\frac{V_{DS}}{2}\right)\frac{V_{DS}}{\left(1+\frac{V_{DS}}{V_{sat}}\right)}$$

$V_{sat}=E_{sat}L$

V_{sat} smaller for shorter FETs

Current Saturation Via Velocity Saturation (A Short Channel Effect)

In the linear region:

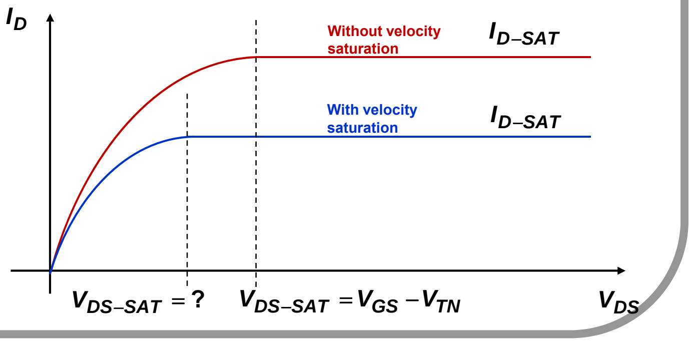
$$I_D = \frac{W}{L}\mu_n C_{ox}\left(V_{GS} - V_{TN} - \frac{V_{DS}}{2}\right)\frac{V_{DS}}{\left(1 + \dfrac{V_{DS}}{V_{sat}}\right)} \qquad \left\{ V_{sat} = E_{sat}L \right.$$

If V_{sat} was ∞, the above expression, as a function of V_{DS}, has a maximum when V_{DS} equals $V_{GS} - V_{TN}$ (corresponding to channel pinch-off near the drain end)

But if V_{sat} was very small, much smaller than $V_{GS} - V_{TN}$, the above expression has a maximum when V_{DS} is smaller than $V_{GS} - V_{TN}$

$$V_{sat} = E_{sat}L$$

V_{sat} is smaller for smaller channel length devices



Current Saturation Via Velocity Saturation (A Short Channel Effect)

In the linear region:

$$I_D = \frac{W}{L}\mu_n C_{ox}\left(V_{GS} - V_{TN} - \frac{V_{DS}}{2}\right)\frac{V_{DS}}{\left(1 + \dfrac{V_{DS}}{V_{sat}}\right)} \qquad \left\{ V_{sat} = E_{sat}L \right.$$

In saturation (due to velocity saturation):

$$\frac{dI_D}{dV_{DS}} = 0 \longrightarrow V_{DS-SAT} = \frac{2(V_{GS} - V_{TN})}{1 + \sqrt{1 + \dfrac{2(V_{GS} - V_{TN})}{V_{sat}}}} < (V_{GS} - V_{TN})$$

The expression for $I_{DS\text{-}SAT}$ is complicated but in the limit $V_{sat} << V_{GS}\text{-}V_{TN}$ (extreme short channel FET):

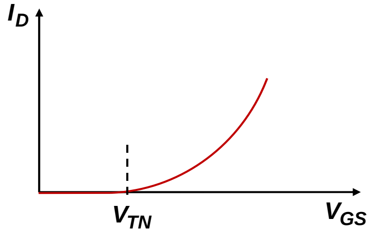
$$\begin{aligned} I_{D-SAT} &= \frac{W}{L}\mu_n C_{ox}(V_{GS} - V_{TN})V_{sat} \\ &= \frac{W}{L}\mu_n C_{ox}(V_{GS} - V_{TN})E_{sat}L \\ &= W\,(\mu_n E_{sat})C_{ox}(V_{GS} - V_{TN}) \\ &= W\,v_{dn-sat}\,C_{ox}(V_{GS} - V_{TN}) \end{aligned} \longrightarrow$$

Note the linear dependence of the saturation current on the gate-to-source voltage (this is a characteristic behavior of velocity saturation)

Current Saturation Via Velocity Saturation
(A Short Channel Effect)

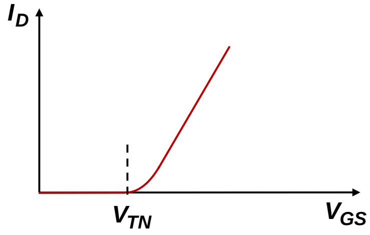
A long channel FET (current saturation via channel pinch-off):

$$I_{D-SAT} \approx \frac{k_n}{2}\left(V_{GS} - V_{TN}\right)^2$$



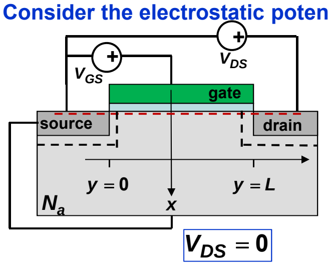
A very short channel FET (current saturation via velocity saturation):

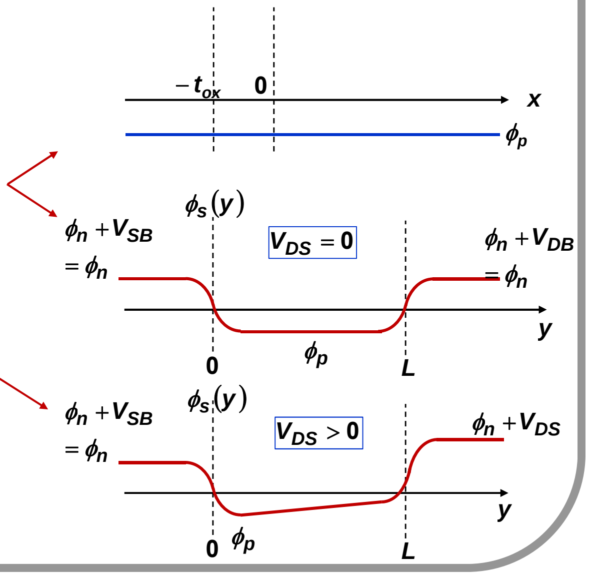
$$I_{D-SAT} = W\, v_{dn-sat}\, C_{ox}\left(V_{GS} - V_{TN}\right)$$



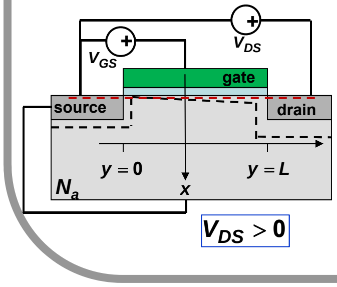
Electrostatic Potential in FETs below Threshold

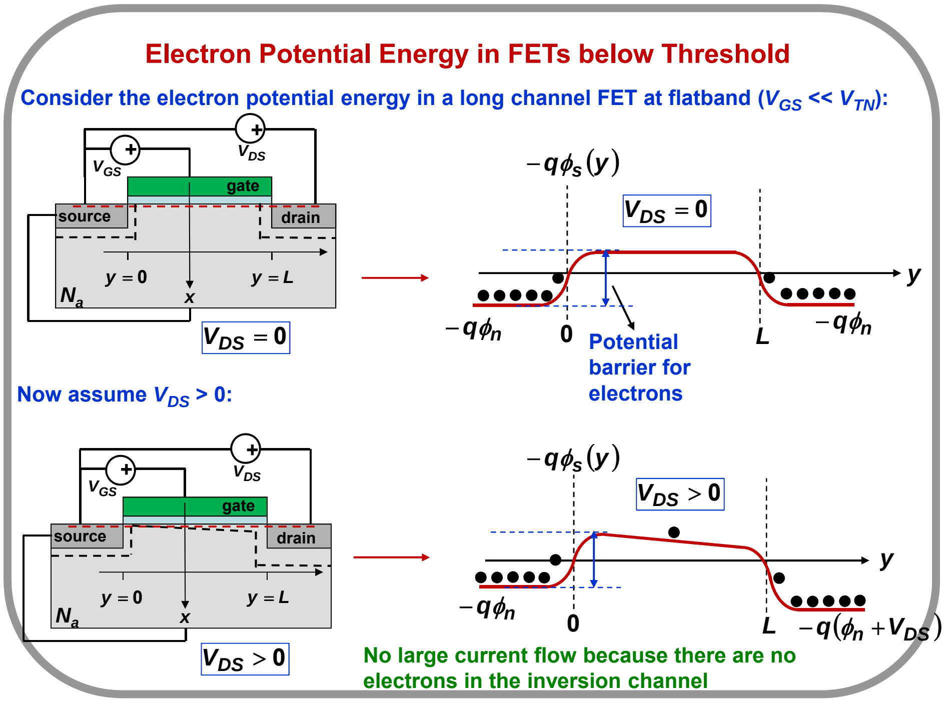
Consider the electrostatic potential in a long channel FET at flatband (V_{GS} << V_{TN}):





Now assume V_{DS} > 0:



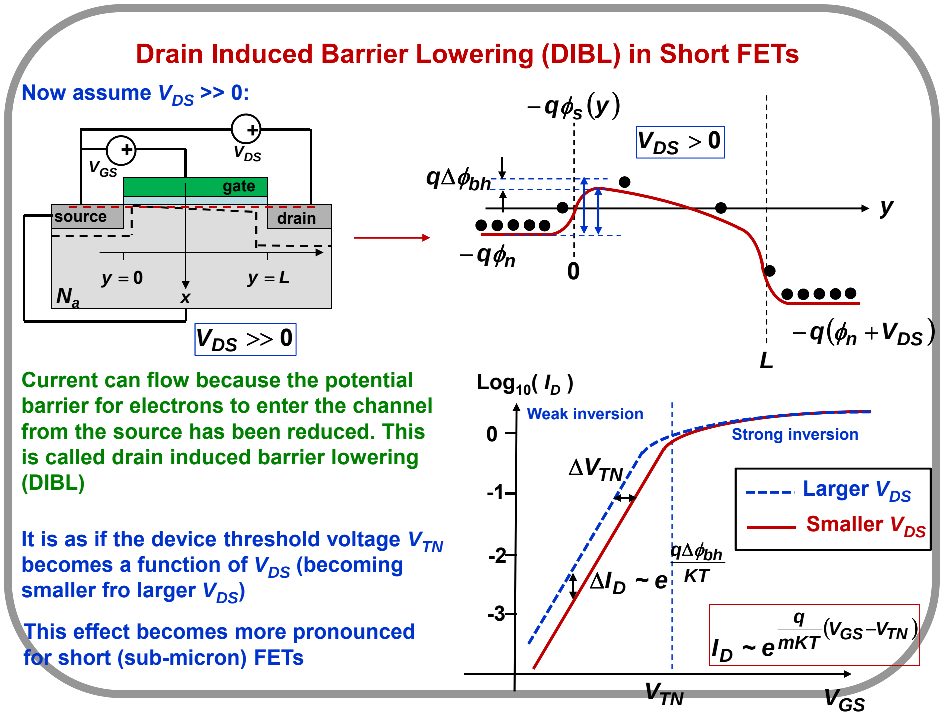


Electron Potential Energy in FETs below Threshold

Consider the electron potential energy in a long channel FET at flatband ($V_{GS} << V_{TN}$):

Now assume $V_{DS} > 0$:

No large current flow because there are no electrons in the inversion channel



Drain Induced Barrier Lowering (DIBL) in Short FETs

Now assume $V_{DS} >> 0$:

Current can flow because the potential barrier for electrons to enter the channel from the source has been reduced. This is called drain induced barrier lowering (DIBL)

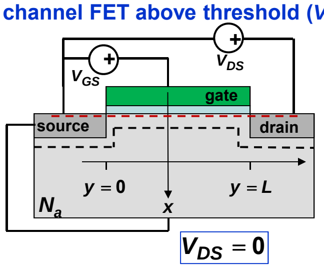
It is as if the device threshold voltage V_{TN} becomes a function of V_{DS} (becoming smaller fro larger V_{DS})

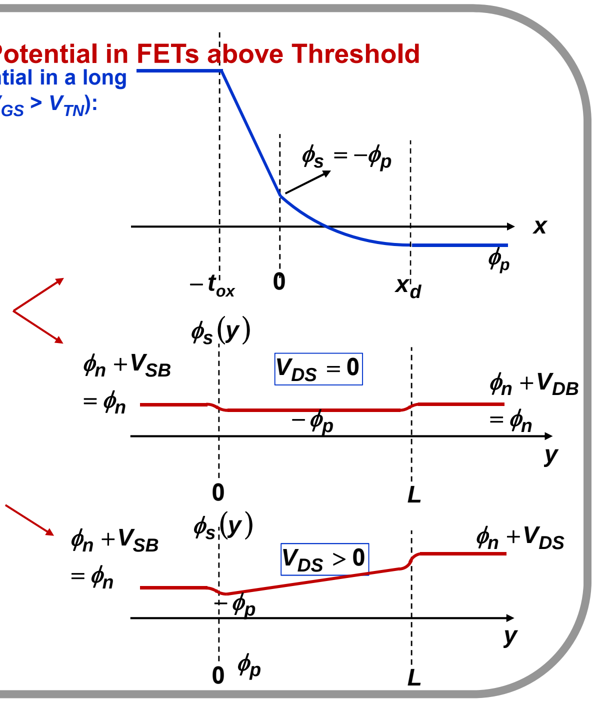
This effect becomes more pronounced for short (sub-micron) FETs

$$I_D \sim e^{\frac{q}{mKT}(V_{GS}-V_{TN})}$$

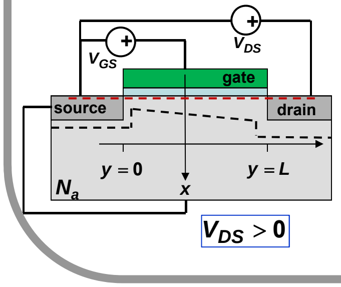
Electrostatic Potential in FETs above Threshold

Consider the electrostatic potential in a long channel FET above threshold (V_{GS} > V_{TN}):



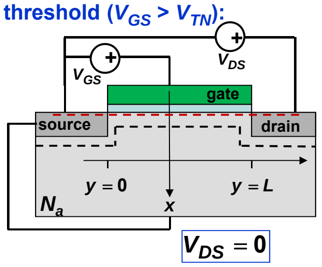


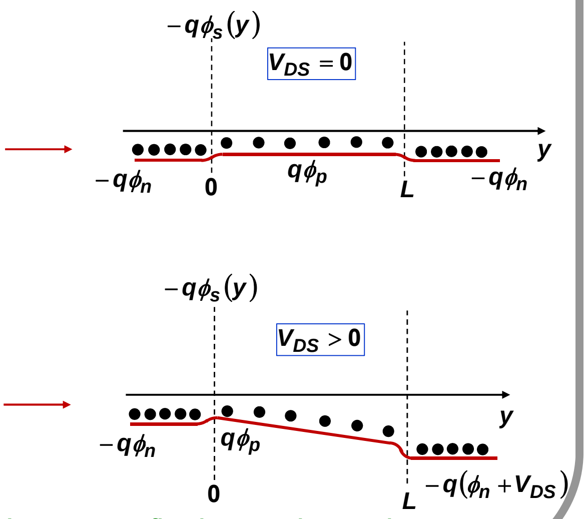
Now assume V_{DS} > 0:



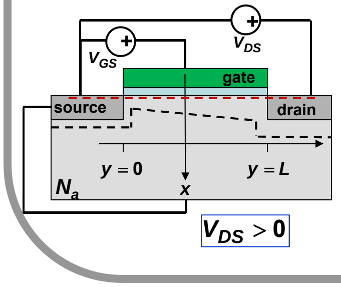
Electron Potential Energy in FETs above Threshold

Consider the electron potential energy in a long channel FET above threshold (V_{GS} > V_{TN}):





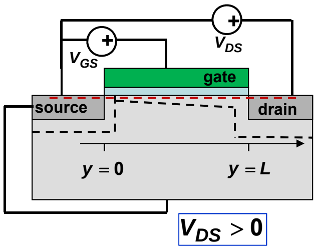
Now assume V_{DS} > 0:

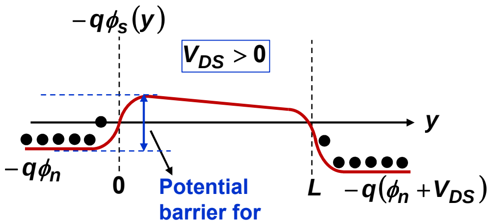


Large current flow because there are lots of electrons in the inversion channel

Punch-Through in FETs (A Short Channel Effect)

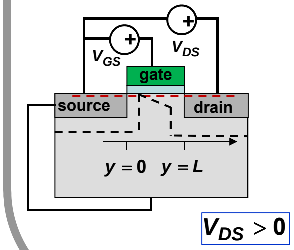
Consider the electron potential energy in a long channel FET below threshold (V_{GS} < V_{TN}):

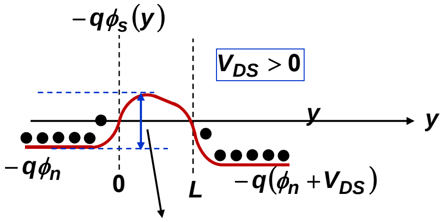




Potential barrier for electrons

Now make the channel length shorter (V_{GS} < V_{TN}):

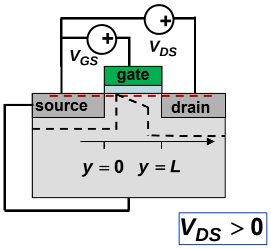


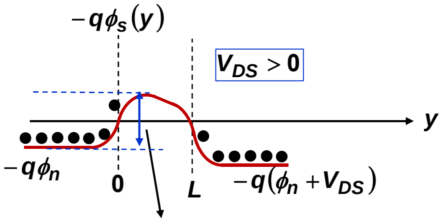


Potential barrier for electrons is still there

Punch-Through in FETs (A Short Channel Effect)

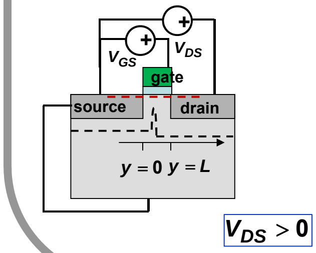
Channel is short (V_{GS} < V_{TN}):

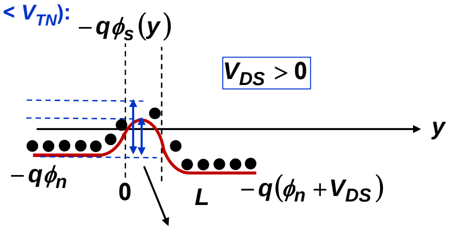




Potential barrier for electrons

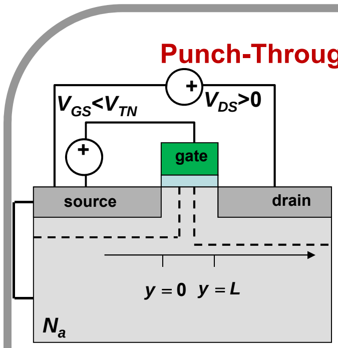
Now make the channel even shorter (V_{GS} < V_{TN}):

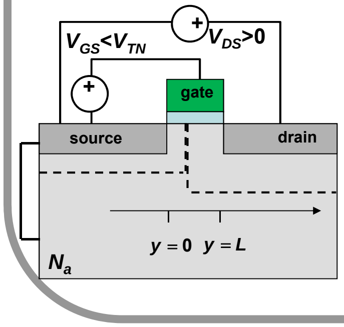




Potential barrier for electrons is significantly reduced when the source-drain depletion regions overlap (this is "punch-through")

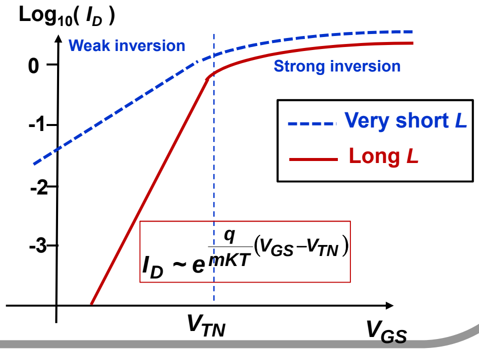
Punch-Through in FETs (A Short Channel Effect)





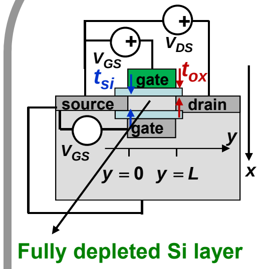
Consider a nano-scale FET with $V_{GS} < V_{TN}$ but $V_{DS} > 0$

- When the depletion regions of the source and drain come close, the barrier to electron flow from the source to the drain reduces significantly
- **This results in leakage current even when $V_{GS} < V_{TN}$!!**
- The problem is particularly bad when V_{DS} is large
- The leakage current contributes to **subthreshold** conduction
- **Therefore, it is difficult to completely turn-off nano-scale FETs!**



$$I_D \sim e^{\frac{q}{mKT}(V_{GS}-V_{TN})}$$

Mitigating Short Channel Effects: The Double-Gate FET

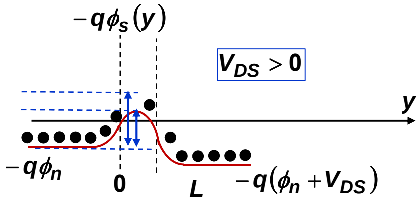


To understand why double-gate FET mitigates the short channel effects one needs to start from the 2D Poisson equation:

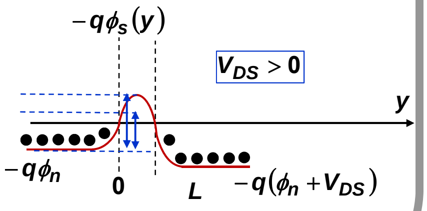
$$\frac{\partial^2\phi(x,y)}{\partial x^2}+\frac{\partial^2\phi(x,y)}{\partial y^2}=\frac{qN_a}{\varepsilon_s}$$

If the curvature of the potential in the semiconductor in the x-direction is reduced, the curvature in the y-direction will increase (to satisfy Poisson equation)

Helps control DIBL and punch-through in nano-scale FETs

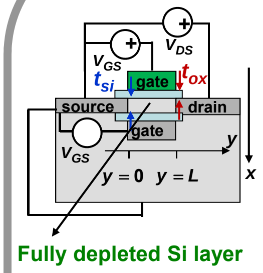


Single-Gate Device



Double-Gate Device

Mitigating Short Channel Effects: The Double-Gate FET



To understand why double-gate FET mitigates the short channel effects one needs to start from the 2D Poisson equation:

$$\frac{\partial^2\phi(x,y)}{\partial x^2}+\frac{\partial^2\phi(x,y)}{\partial y^2}=\frac{qN_a}{\varepsilon_s}$$

The main result from 2D Poisson equation:

In order to reduce the short channel effects, minimize the value of the length parameter:

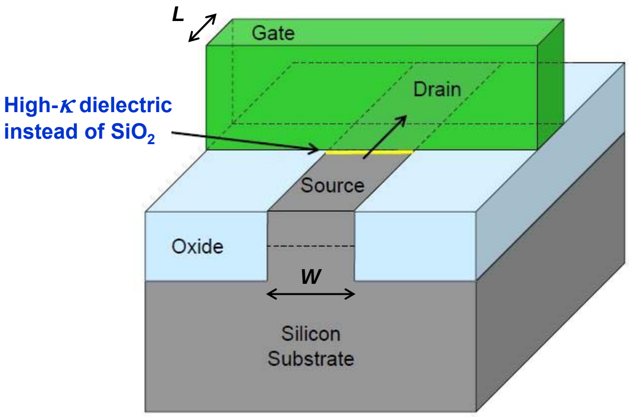
$$\sqrt{\frac{\varepsilon_s}{\varepsilon_{ox}}t_{ox}t_{Si}}$$

This implies:

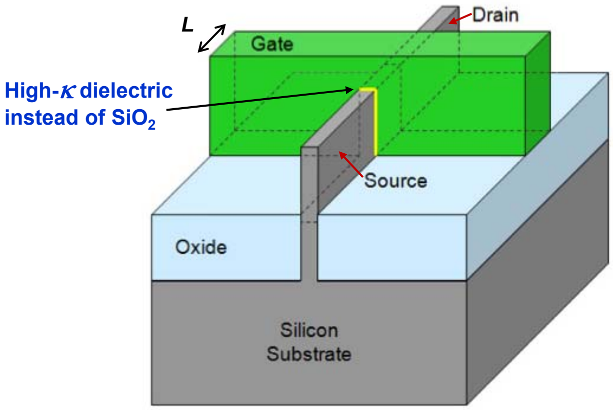
- Choose a high-dielectric-constant gate dielectric (in place of SiO_2) (e.g. use HfO_2)
- Keep the thickness t_{ox} of the gate dielectric layer as small as possible
- Keep the thickness t_{Si} of the Si layer as small as possible

The Traditional Planar Single-Gate FET

$$\boxed{\varepsilon_{ox} = \varepsilon_o\kappa}$$



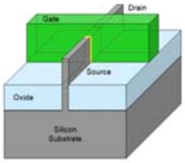
The Fin-FET: Gate All-Around 3D Device Geometry

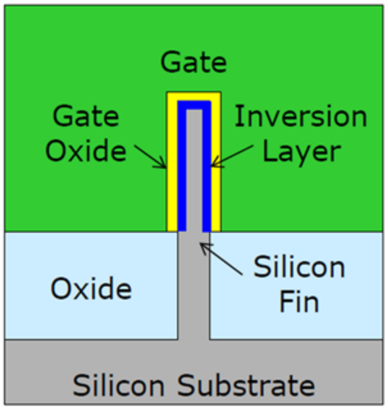


Technology of choice for the sub-30 nm FETs

Smaller footprint
Better control over short channel effects

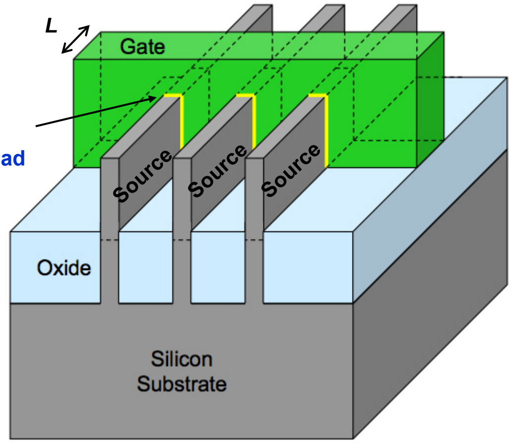
The FinFET





Side View

The Multiple-Fin FinFET



High–κ dielectric instead of SiO_2

Can use any number of fins per FET for increased current drive!

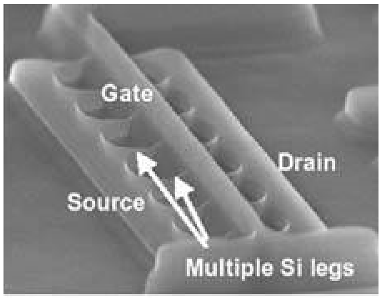
Technology of choice for the sub-30 nm FETs

Superior current drive

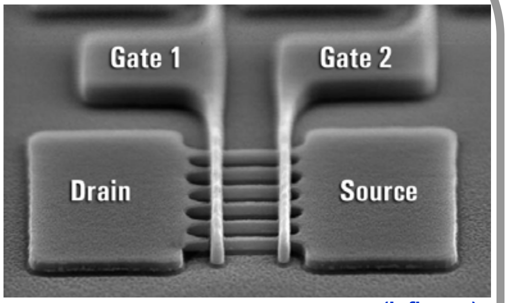
Smaller footprint

Better control over short channel effects

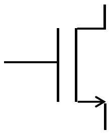
The Multiple-Fin FinFETs



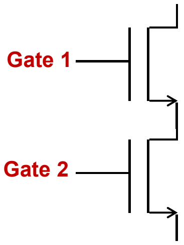
(Intel)



(Infineon)

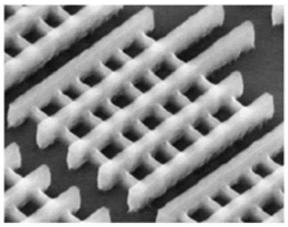


One FET

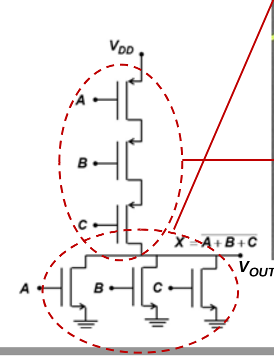


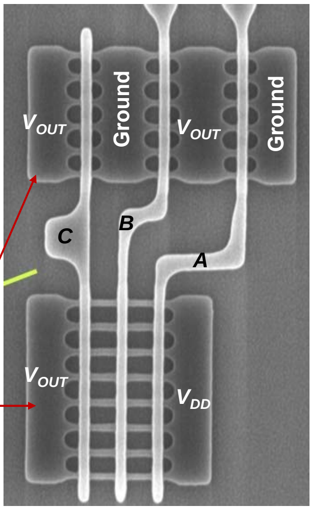
Two FETs in series

Multiple-Fin FinFET Logic Gates



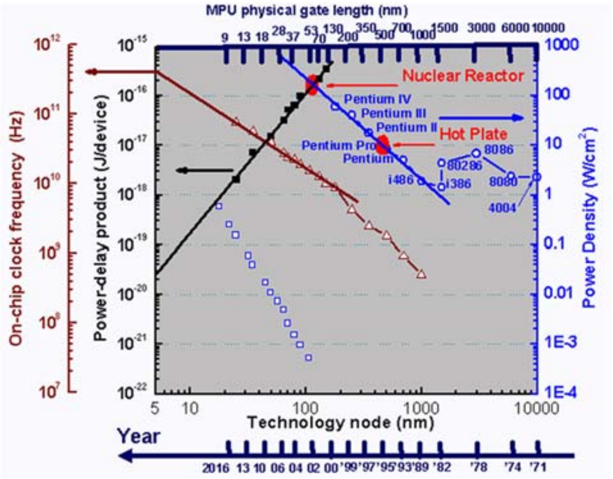
Intel





(IMEC: Three input NOR gate using Fin-FETs)

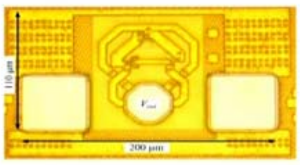
Technology Trends

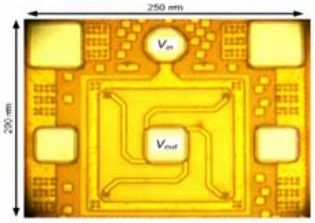


Proceedings of the IEEE, 96, 212 (2008)

Technology miniaturization allows for higher integration density and faster switching speeds, and the power-delay product for a single device continues to decrease.

Making CMOS Analog Circuits go Terahertz (100-1000 GHz)

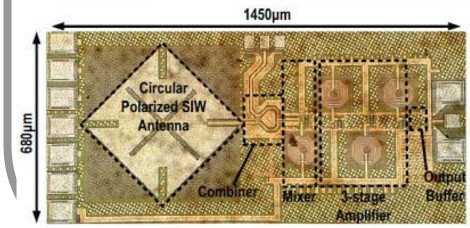




Prof. Afshari's group at Cornell

An oscillator with 0.16mW power at 482 GHz in 65 nm CMOS

A 220-275 GHz frequency doubler with 0.22mW power at 244 GHz in 65 nm CMOS



A ~281 GHz CMOS RF imaging chip