Chapter 7 MOSFET Technology Scaling, Leakage Current and Other Topics

7.1 Technology Scaling – Small is Beautiful

YEAR	1992	1995	1997	1999	2001	2004	2007	2010
Technology Generation	0.5 μm	0.35 μm	0.25 μm	0.18 μm	0.13 μm	90 nm	65 nm	45 nm

- New technology node every three years or so. Defined by minimum metal line width.
- All feature sizes, e.g. gate length, are ~70% of previous node.
- **Reduction of circuit size by 2 – good for cost.**

International Technology Roadmap for Semiconductors, 1999 Edition

Year of Shipment	**1999**	**2002**	**2005**	**2008**	**2011**	**2014**
DRAM metal half pitch (nm)	180	130	100	70	50	35
MPU physical L_g (nm)	140	85	65	45	32	22
T_{ox}(nm)	1.5-1.8	1.5-1.9	1-1.5	0.8-1.2	0.6-0.8	0.5-0.6
V_{DD}	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6	0.3-0.6
$I_{on,HP}$ (μA/μm)	750/350	750/350	750/350	750/350	750/350	750/350
$I_{off,HP}$ (nA/μm)	5	10	20	40	60	160
$I_{on,LP}$ (μA/μm)	490/230	490/230	490/230	490/230	490/230	490/230
$I_{off,LP}$ (pA/μm)	7	10	20	40	80	160

→ No known solutions

- V_{dd} is reduced at each node to contain power consumption in spite of rising transistor density and frequency
- T_{ox} is reduced to raise I_{on} for speed consideration

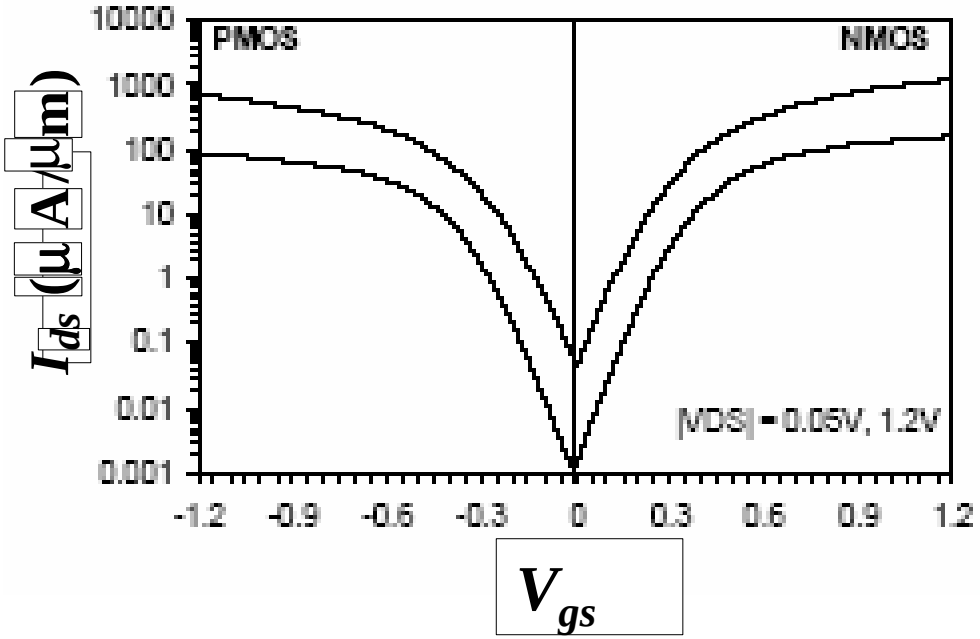
Excerpt of 2003 ITRS Scaling to 2016

Year of Production	2004	2007	2010	2013	2016
Technology Node (nm)	90	65	45	32	22
HP physical Lg (nm)	37	25	18	13	9
EOT(nm) (HP/LSTP)	1.2/2.1	0.9/1.6	0.7/1.3	0.6/1.1	0.5/1.0
Vdd (HP/LSTP)	1.2/1.2	1.1/1.1	1.1/1.0	1.0/0.9	0.9/0.8
Ion/W,HP (mA/mm)	1100	1510	1900	2050	2400
Ioff/W,HP (mA/mm)	0.05	0.07	0.1	0.3	0.5
Ion/W,LSTP (mA/mm)	440	510	760	880	860
Ioff/W,LSTP (mA/mm)	1e-5	1e-5	6e-5	8e-5	1e-4

- HP: High Performance Logic
- LSTP: Low Standby Power Technology

7.2 Subthreshold Current

- The leakage current that flows at $V_g<V_t$ is called the subthreshold current.

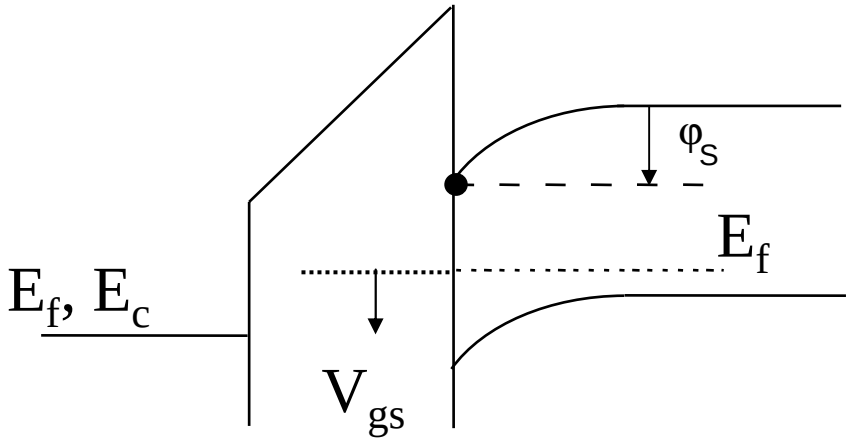


Intel, T. Ghani et al., IEDM 2003

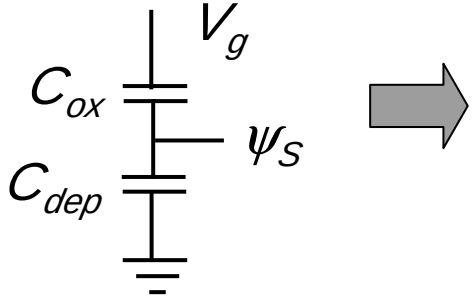
90nm technology.
Gate length: 45nm for NMOS, 50nm for PMOS

- The current at $V_{gs}=0$ and $V_{ds}=V_{dd}$ is called $I_{off.}$

- Subthreshold current $\propto n_s$ (surface inversion carrier concentration)
- $n_s \propto e^{q\varphi s/kT}$



- φ_s varies with V_g through a capacitor network

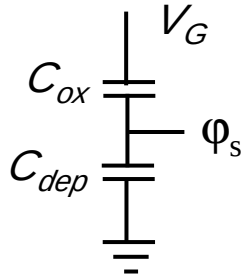


$$\frac{d\varphi_s}{dV_g} = \frac{C_{oxe}}{C_{oxe} + C_{dep}} = \frac{1}{\eta}$$

In subthreshold, $\varphi_s = constant + V_g/\eta$

Subthreshold Leakage Current

$$I_{ds} \propto n_s \propto e^{q\varphi_s / kT} \propto e^{q\left(\text{constant} + V_{gs}/\eta\right)/kT} \propto e^{qV_{gs}/\eta kT}$$



$$I_{ds} \propto e^{qV_{gs}/\eta kT}$$

$$\eta = 1 + \frac{C_{dep}}{C_{oxe}}$$

- Subthreshold current changes 10x for $\eta \cdot 60mV$ change in V_g.

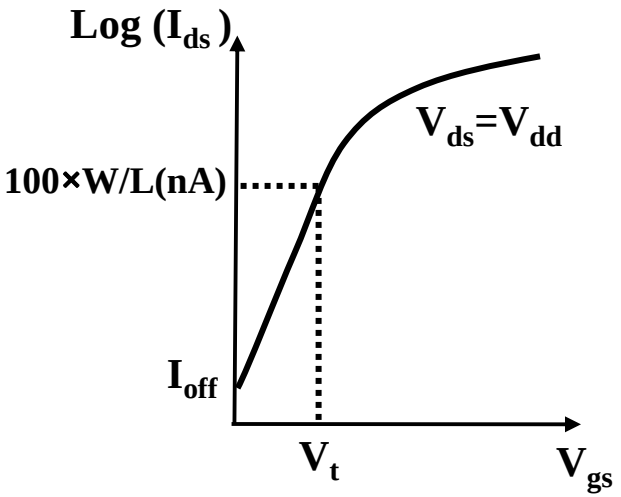
Reminder: 60mV is (ln10)·kT/q

•Subthreshold swing, S : the change in V_{gs} corresponding to 10x change in subthreshold current. $S = \eta \cdot 60mV$, typically 80-100mV

Subthreshold Leakage Current

- Practical definition of V_t : the V_{gs} at which I_{ds}= 100nA×W/L

=> $$I_{subthreshold}(nA) \approx 100 \times \frac{W}{L} \times e^{q(V_g - V_t)/\eta kT} = 100 \times \frac{W}{L} \times 10^{(V_g - V_t)/S}$$



$$I_{off}(nA) = 100 \times \frac{W}{L} \times 10^{-V_t/S}$$

is determined only by V_t and subthreshold swing.

Subthreshold Swing (S)

- Smaller S is desirable (lower I_{off} for a given V_t). Minimum possible value of S is 60mV/dec.

- How do we lower swing?

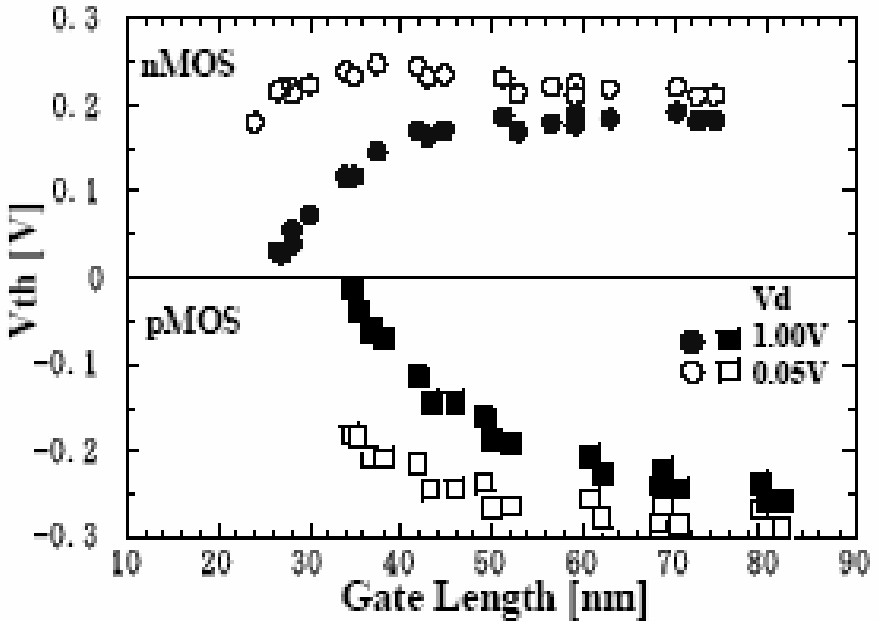
$$S = 60mV \cdot \left(1 + \frac{C_{dep}}{C_{oxe}}\right)$$

  - Thinner T_{ox} => larger C_{oxe}
  - Lower substrate doping => smaller C_{dep}
  - Lower temperature

- Limitations
  - Thinner T_{ox} – oxide breakdown reliability or oxide leakage current
  - Lower substrate doping – doping is not a free parameter but set by V_t.

7.3 V_t Roll-off

- V_t roll-off: V_t decreases with decreasing L_g.
- It determines the minimum acceptable L_g because I_{off} is too large if V_t becomes too small.

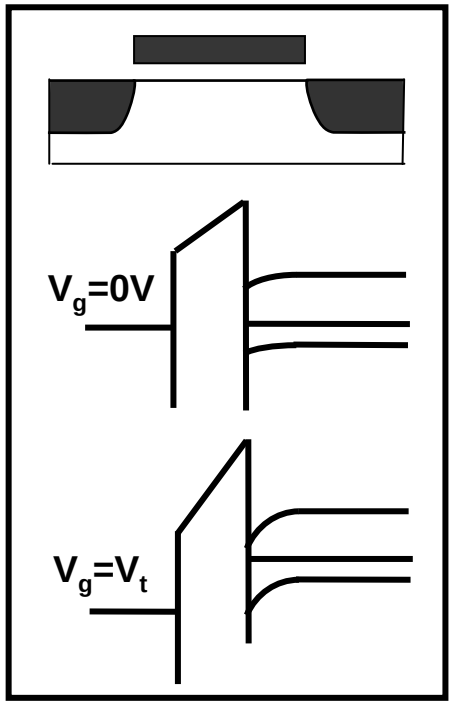


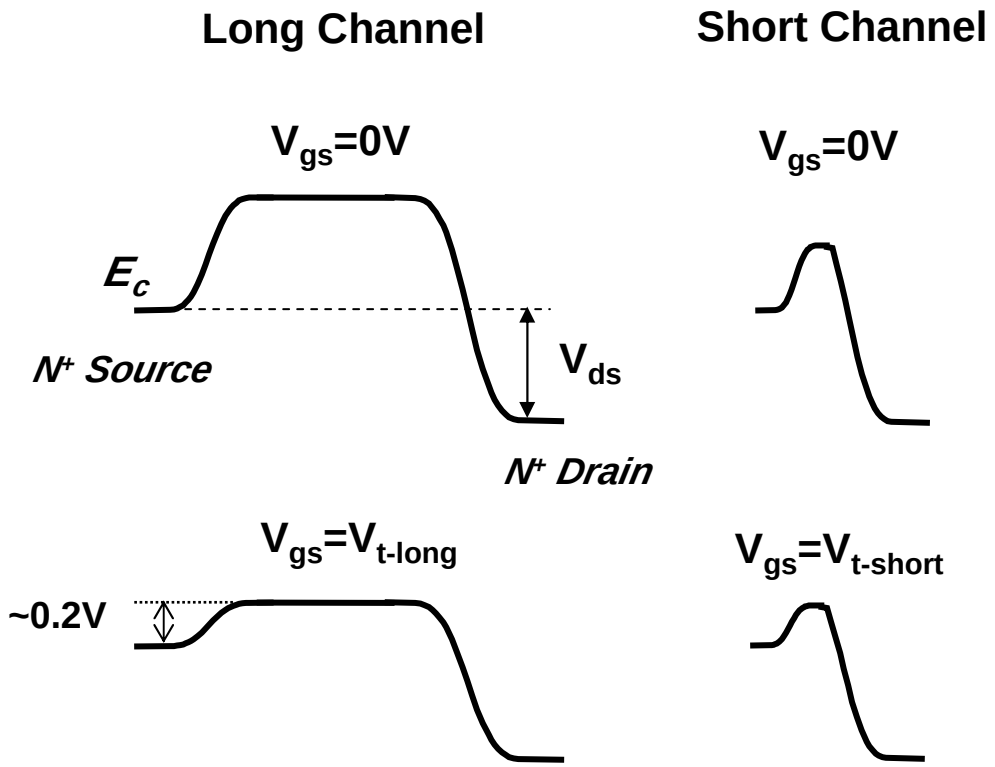
K. Goto et al., (Fujitsu) IEDM 2003 65nm technology. EOT=1.2nm, V_{dd}=1V

- ***Question: Why data is plotted against L_g, not L?***

  Answer: L is difficult to measure. L_g is. Also, L_g is the quantity that manufacturing engineers can control directly.

Why Does V_t Decrease with L? – Potential Barrier Concept

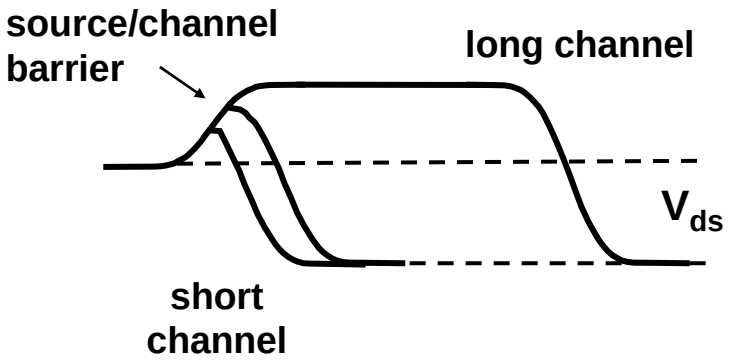




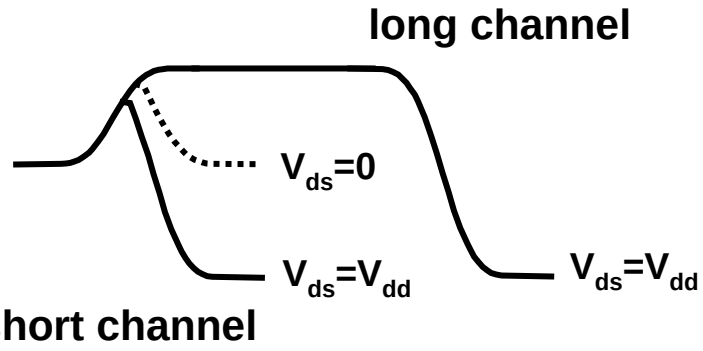
- When L is small, smaller V_g is needed to reduce the barrier to 0.2V, i.e. V_t is smaller.
- V_t roll-off is greater for shorter L

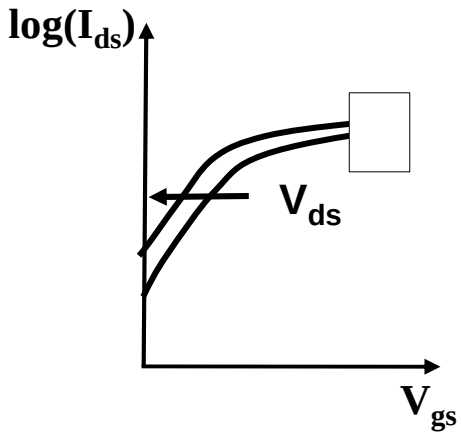
Energy-Band Diagram from Source to Drain

- L dependence

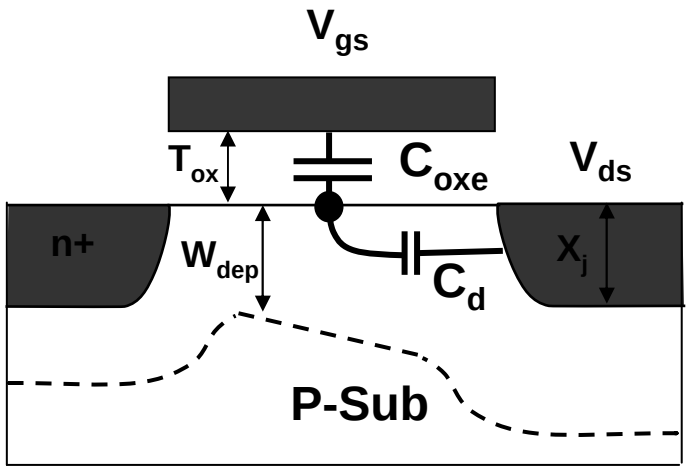


- V_{ds} dependence





V_t Roll-off -- Capacitance Network Model



As the channel length is reduced, drain to channel distance is reduced; therefore C_d increases

V_{ds} helps V_{gs} to invert the surface, therefore

$$V_t = V_{t-long} - V_{ds} \cdot \frac{C_d}{C_{oxe}}$$

$$V_t = V_{t-long} - (V_{ds} + 0.4) \cdot \frac{C_d}{C_{oxe}}$$

Due to built-in potential between N^- channel and N^+ drain & source

- 2-D Poisson Eq. solution shows that C_d is an exponential function of L.

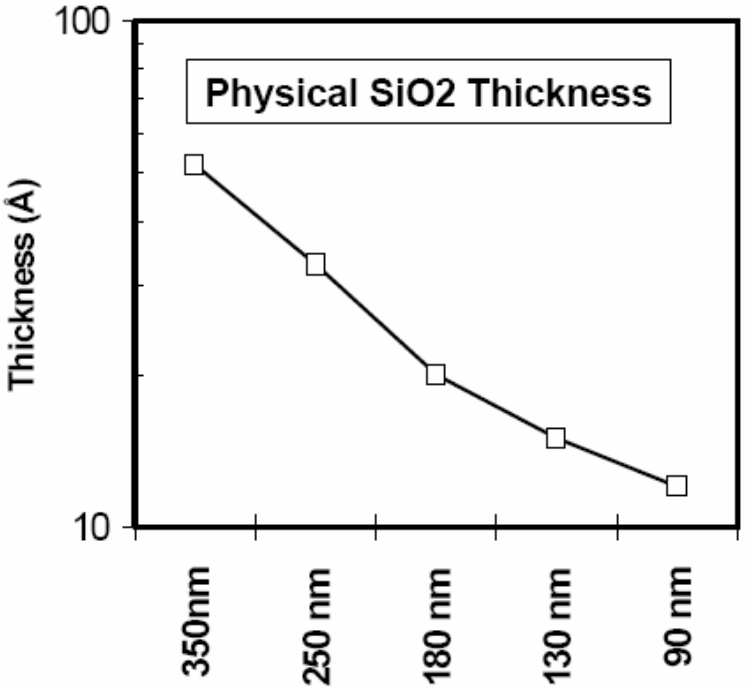
$$V_t = V_{t-long} - (V_{ds} + 0.4) \cdot e^{-L/l_d}$$

where $l_d \approx \sqrt[3]{T_{ox} W_{dep} X_j}$

Vertical dimensions (T_{ox}, W_{dep}, X_j) must be scaled to support L reduction

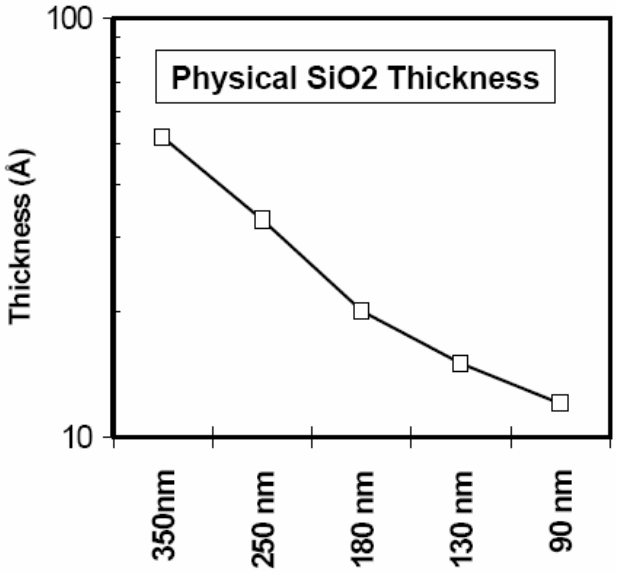
$$V_t = V_{t-long} - (V_{ds} + 0.4) \cdot e^{-L/l_d}$$

where $l_d \approx \sqrt[3]{T_{ox} W_{dep} X_j}$

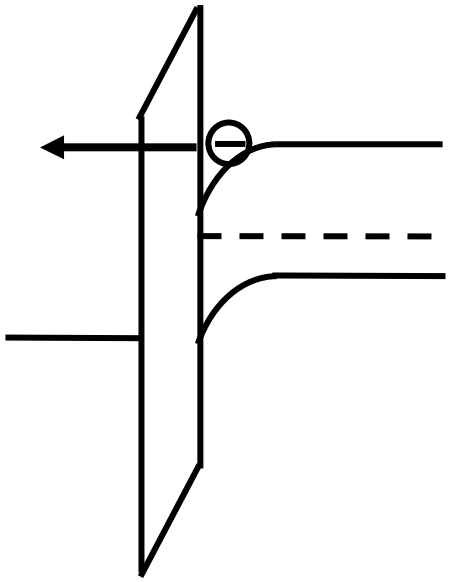


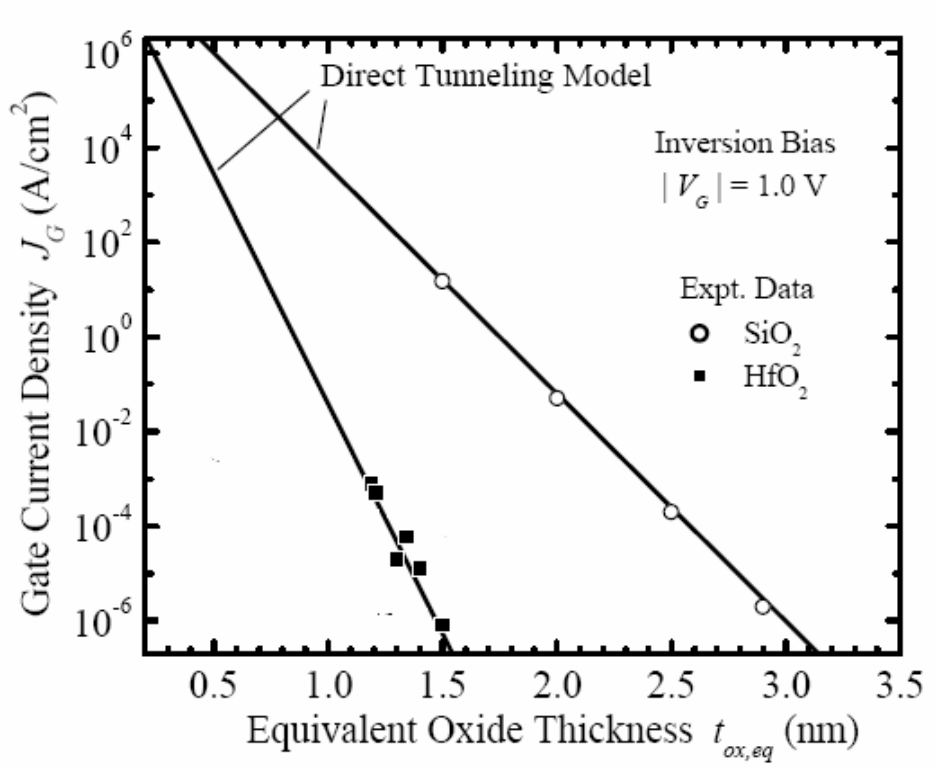
7.4 Reducing the Gate Insulator Thickness and T_{oxe}

- Oxide thickness has been reduced ov the years from 300nm to 1.2nm.
- Why reduce oxide thickness?
  - Larger C_{ox} to raise I_{on}
  - Reduce subthreshold swing
  - Control V_t roll-off
- Thinner is better. However, if the ox is too thin
  - Breakdown due to high electric field
  - Leakage current



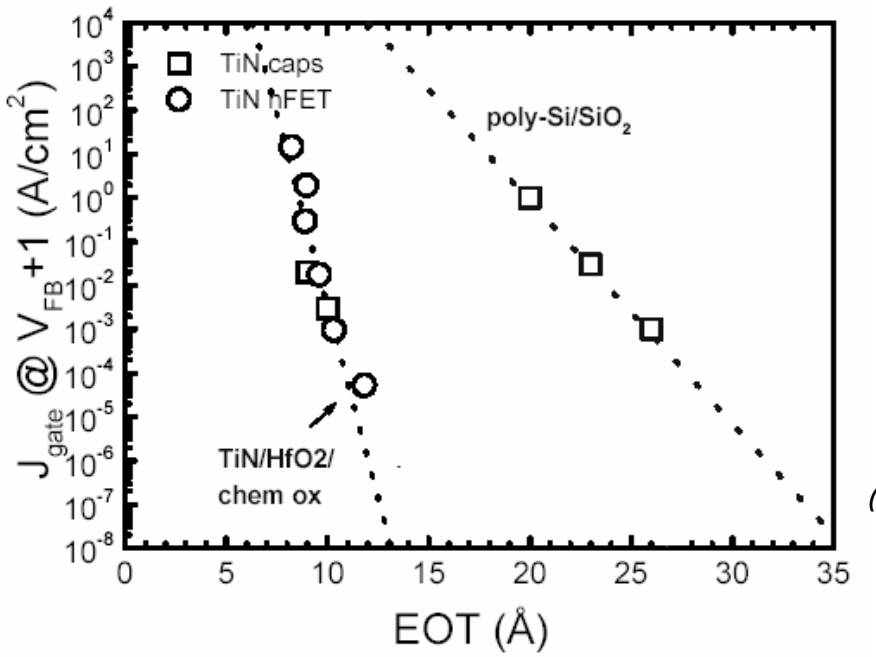
Tunneling Leakage Current





- For SiO_2 films thinner than 1.5nm, tunneling leakage current has become the limiting factor.
- HfO_2 has several orders lower leakage for the same EOT.

Replacing SiO_2 with HfO_2---High-k Dielectric



(After W. Tsai et al., IEDM'03)

- HfO_2 has a relative dielectric constant (k) of ~24, six times large than that of SiO_2.
- For the same EOT, the HfO_2 film presents a much thicker (albeit a lower) tunneling barrier to the electrons and holes.
- Toxe can be further reduced by introducing metal-gate technology since the poly-depletion effect is eliminated.

Challenges of High-K Technology

- The challenges of high-k dielectrics are
  - chemical reactions between them and the silicon substrate and gate,
  - lower surface mobility than the Si/SiO_2 system
  - too low a V_t for P-channel MOSFET (as if there is positive charge in the high-k dielectric).

- A thin SiO_2 interfacial layer may be inserted between Si-substrate and high-k film.

Question: How can T_{inv} be reduced?

(Answer is in Sec. 7.4 text)

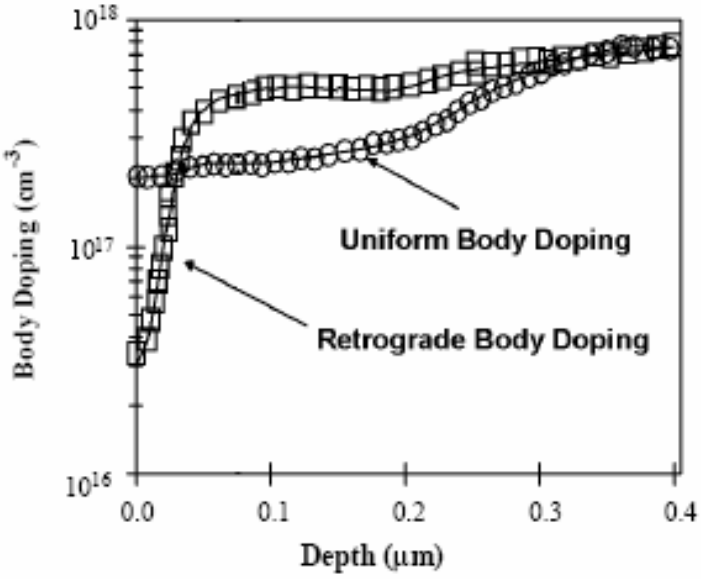
7.5 How to Reduce W_{dep}

- W_{dep} can be reduced by increasing N_{sub}

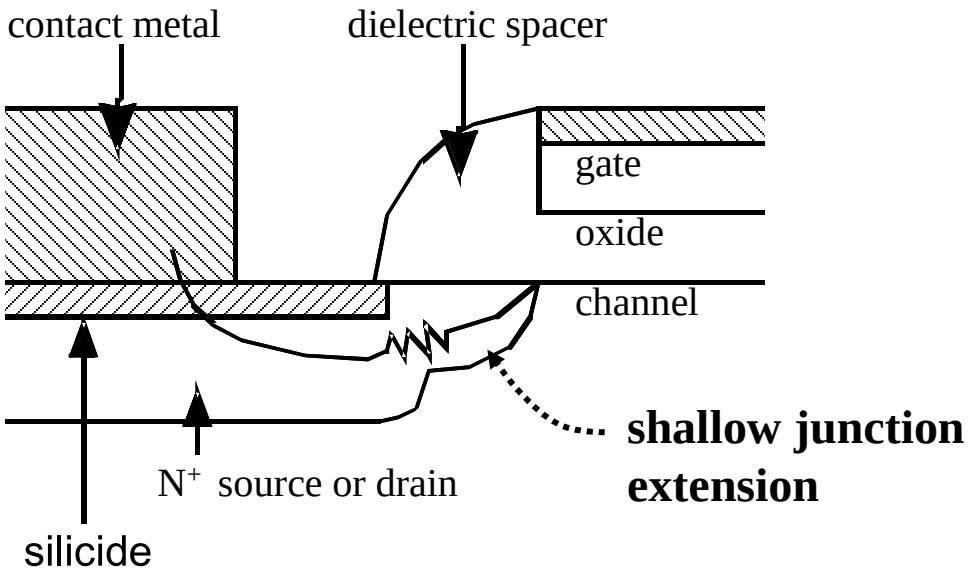
$$V_t = V_{fb} + 2\phi_B + \frac{\sqrt{qN_{sub}2\varepsilon_s 2\phi_B}}{C_{ox}} = V_{fb} + 2\phi_B + \frac{2\varepsilon_s 2\phi_B}{C_{ox}W_{dep}}$$

  - If N_{sub} is increased, C_{ox} should be increased in order to keep V_t the same.
  - W_{dep} can be reduced in proportion to T_{ox}.

- Or use retrograde doping with very thin lightly doped surface layer
  - Also, less impurity scattering in the inversion layer => higher mobility

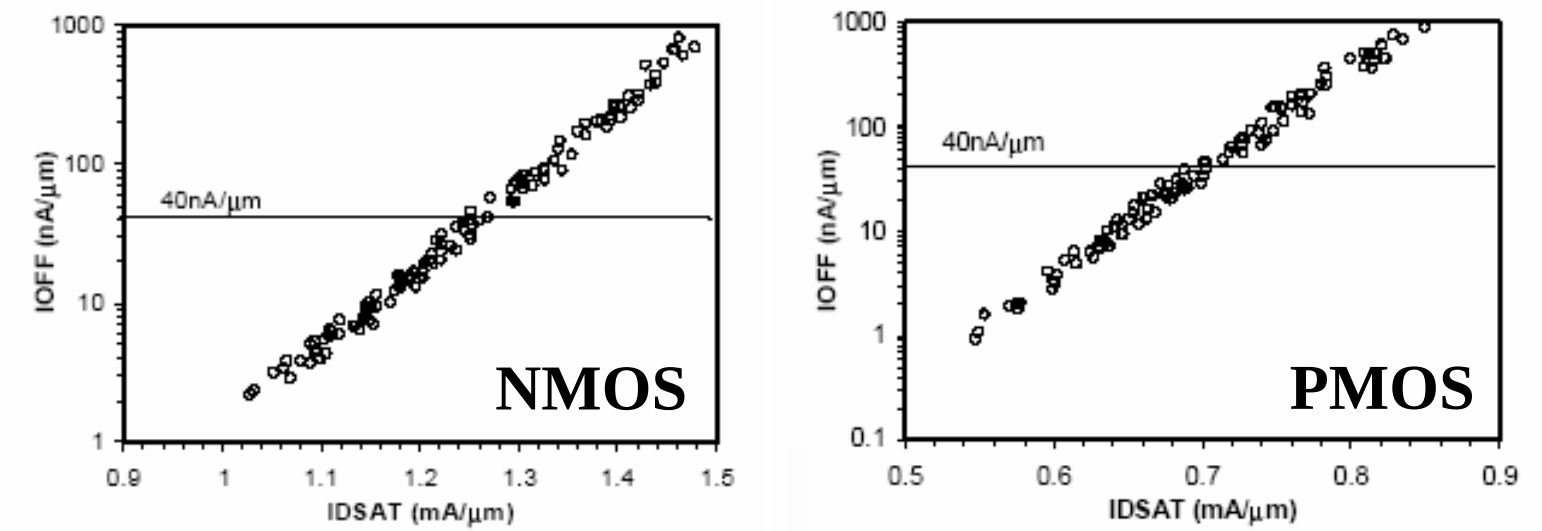


7.6 Shallow Junction Technology



- The shallow junction extension helps to control V_t roll-off.
- Shallow junction and light doping combine to produce an undesirable parasitic resistance that reduces the precious $I_{on.}$
- Theoretically, metal S/D can be used as a very shallow "junction".

7.7 Trade-off between I_{on} and I_{off}



Intel, T. Ghani et al., IEDM 2003

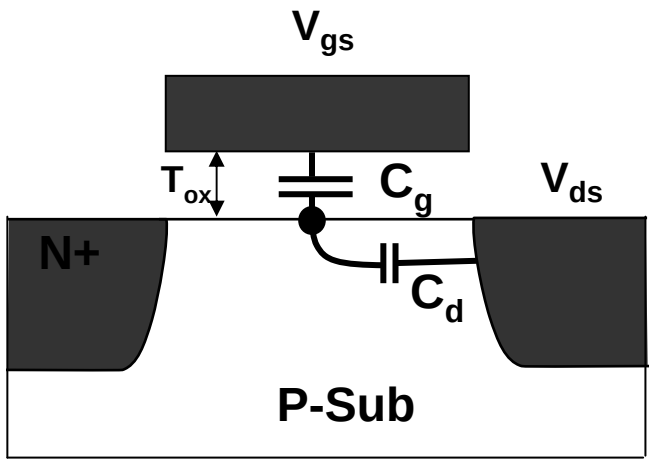
- Higher I_{on} goes hand-in-hand with larger I_{off} -- think L, Vt, Tox, Vdd.
- Figure shows spread in I_{on} (and I_{off}) produced by intentional variation in Lg and unintentional manufacturing variances in L_g and other parameters.

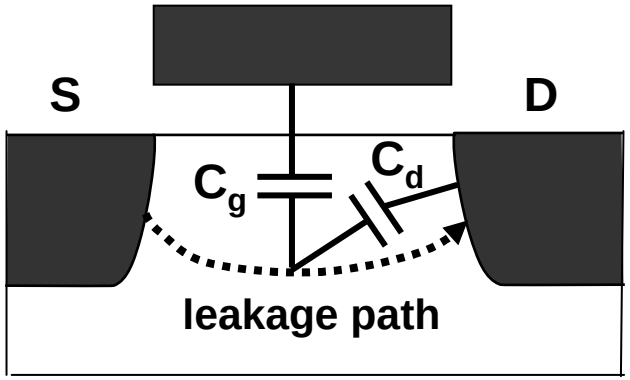
Circuit Techniques to Relax the Conflict between I_{on} and I_{off}

- Substrate (well) bias
  - Only some circuit blocks need to operate at high speed.
  - Can use reverse well bias to raise the V_t for the rest.
  - This techniques can also reduce the chip-to-chip and block-to-block variations with intelligent control circuitry.
- Multiple V_t or V_{dd}
  - Lower V_t or higher V_{dd} are used only in the blocks that need speed
- Alternative MOSFET structures.

7.8 More Scalable Device Structures

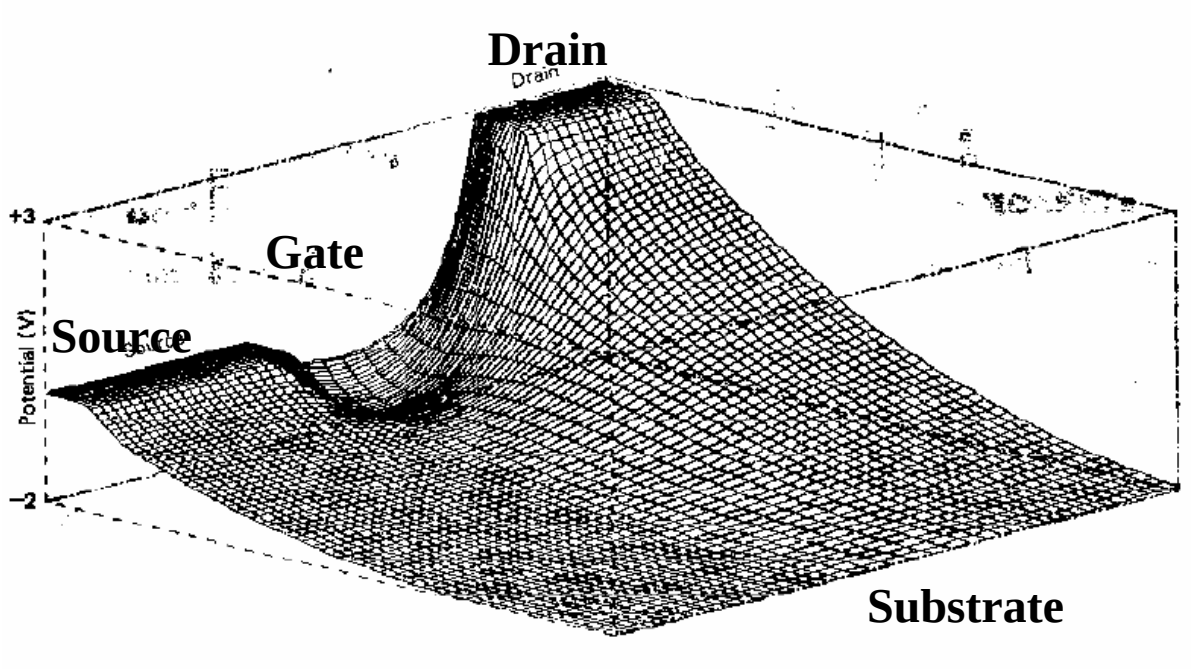
- Vertical Scaling is important. For example, reducing T_{ox} gives the gate excellent control of Si surface potential.
- But, the drain could still have more control than the gate along another leakage current path that is some distance below the Si surface. (Right figure.)





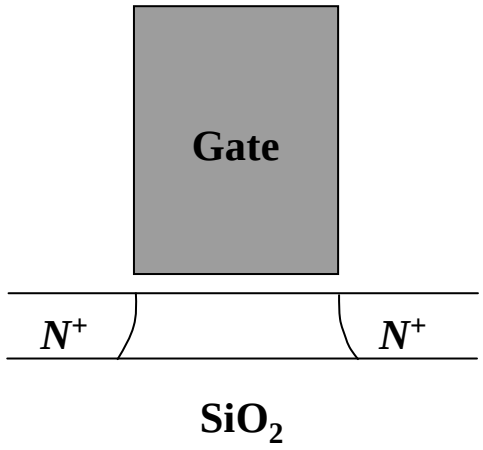
2-D Potential Profile

- The drain voltage can pull the potential barrier down and allow leakage current to flow along a submerged path.

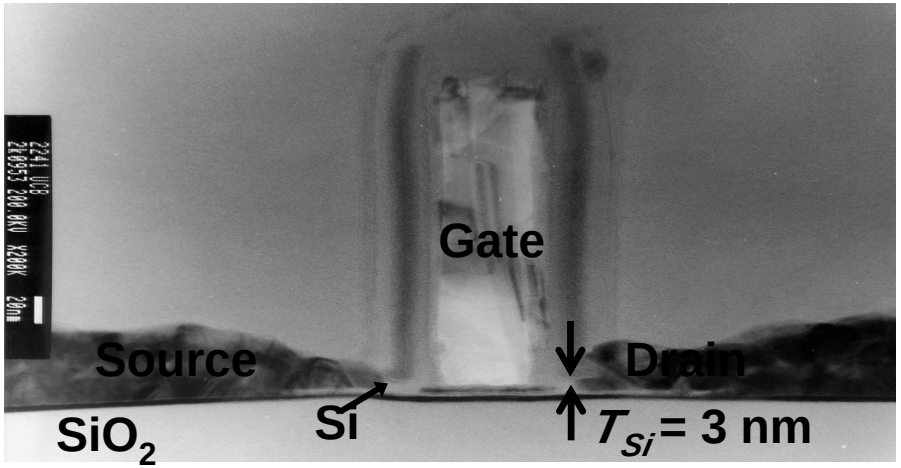


Ultra-Thin-Body (UTB) MOSFET

- MOSFET built on very thin silicon film on an insulator (SiO_2).
- Since the silicon film is very thin, perhaps less than 10nm, no leakage path is very far from the gate.

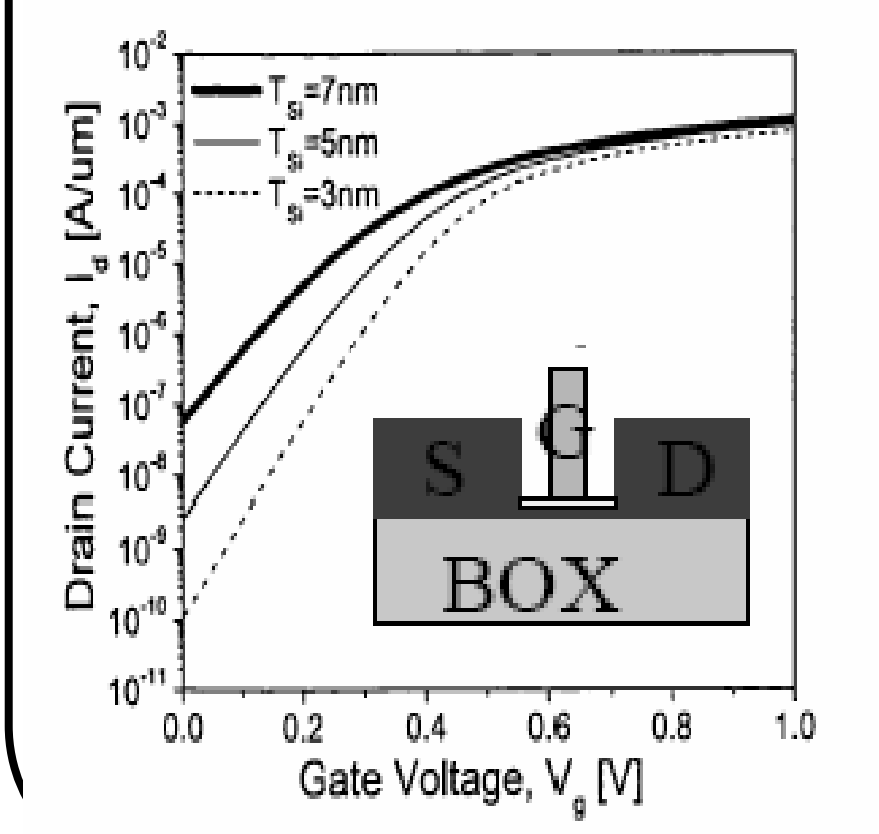


Electron Micrograph of UTB MOSFET

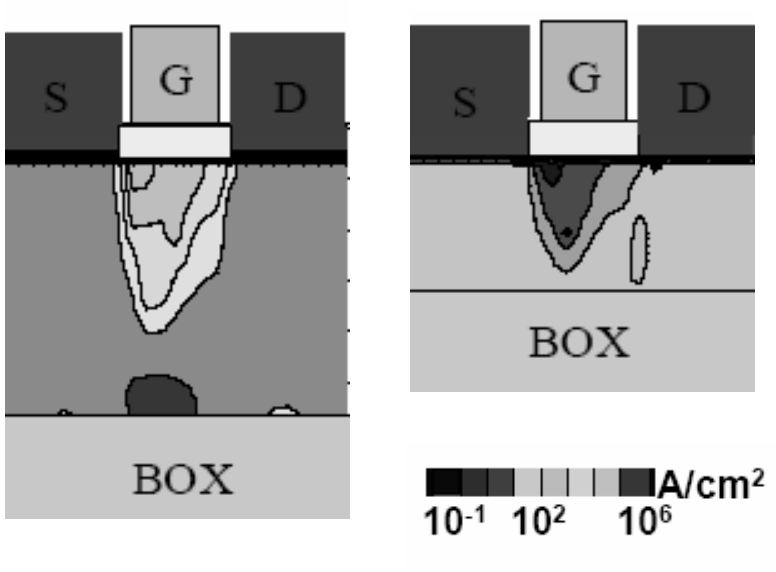


New Structure I--Ultra-Thin-Body MOSFET

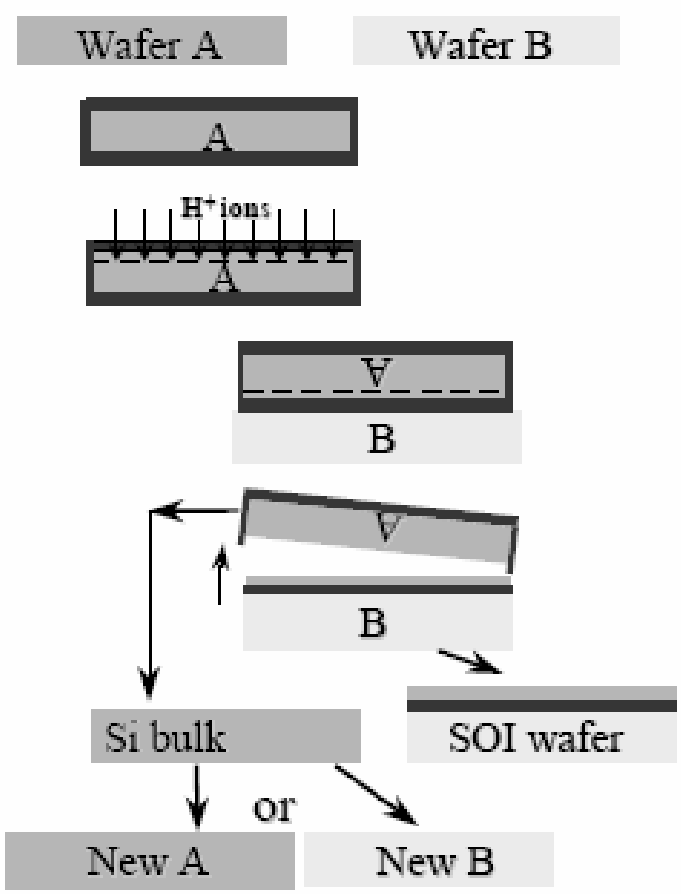
- The subthreshold leakage is reduced as the silicon film is made thinner.



T_{ox}=1.5nm, N_{sub}=1e15cm^{-3}, V_{dd}=1V, V_{gs}=0

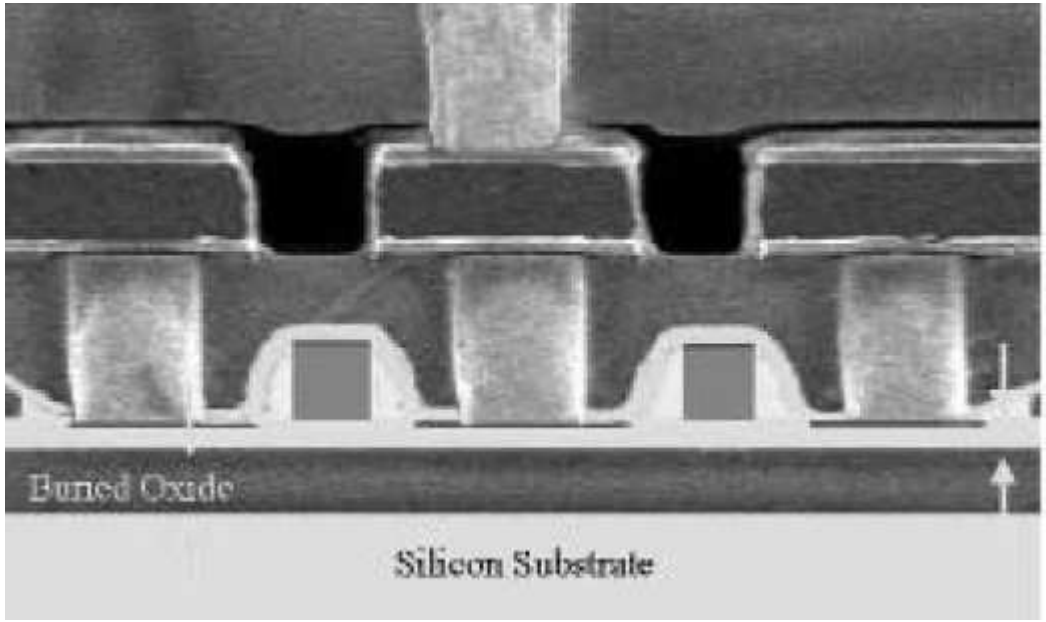


Preparation of Silicon-on-Insulator (SOI) Substrate



- Initial Silicon wafer A and B
- Oxidize wafer A to grow SiO2
- Implant hydrogen into wafer A
- Place wafer A, upside down, over wafer B.
- A low temperature annealing causes the two wafers to fuse together.
- Apply another annealing step to for H_2 bubbles and split wafer A.
- Polish the surface and the SOI wafer is ready for use.
- Wafer A can be reused.

Cross-Section of SOI Circuits

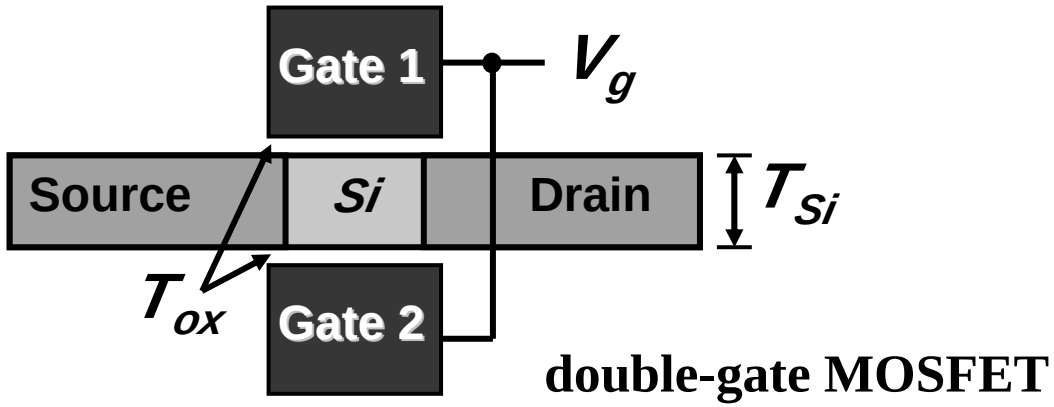


Buried Oxide **Si**

- Due to the high cost of SOI wafers, only some microprocessors, which command high prices and compete on speed, have embraced this technology.
- In order to benefit from the UTB concept, Si film thickness must be agreesively reduced to ~ Lg/4

New Struture II--Multi-Gate MOSFET and FinFET

- The second way of eliminating deep leakage paths is to provide gate control from more than one side of the channel.
- The Si film is very thin so that no leakage path is far from one of the gates.
- Because there are more than one gates, the structure may be called **multi-gate MOSFET**.

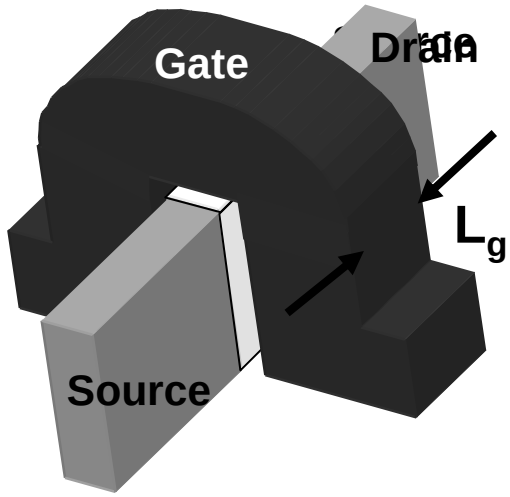


FinFET

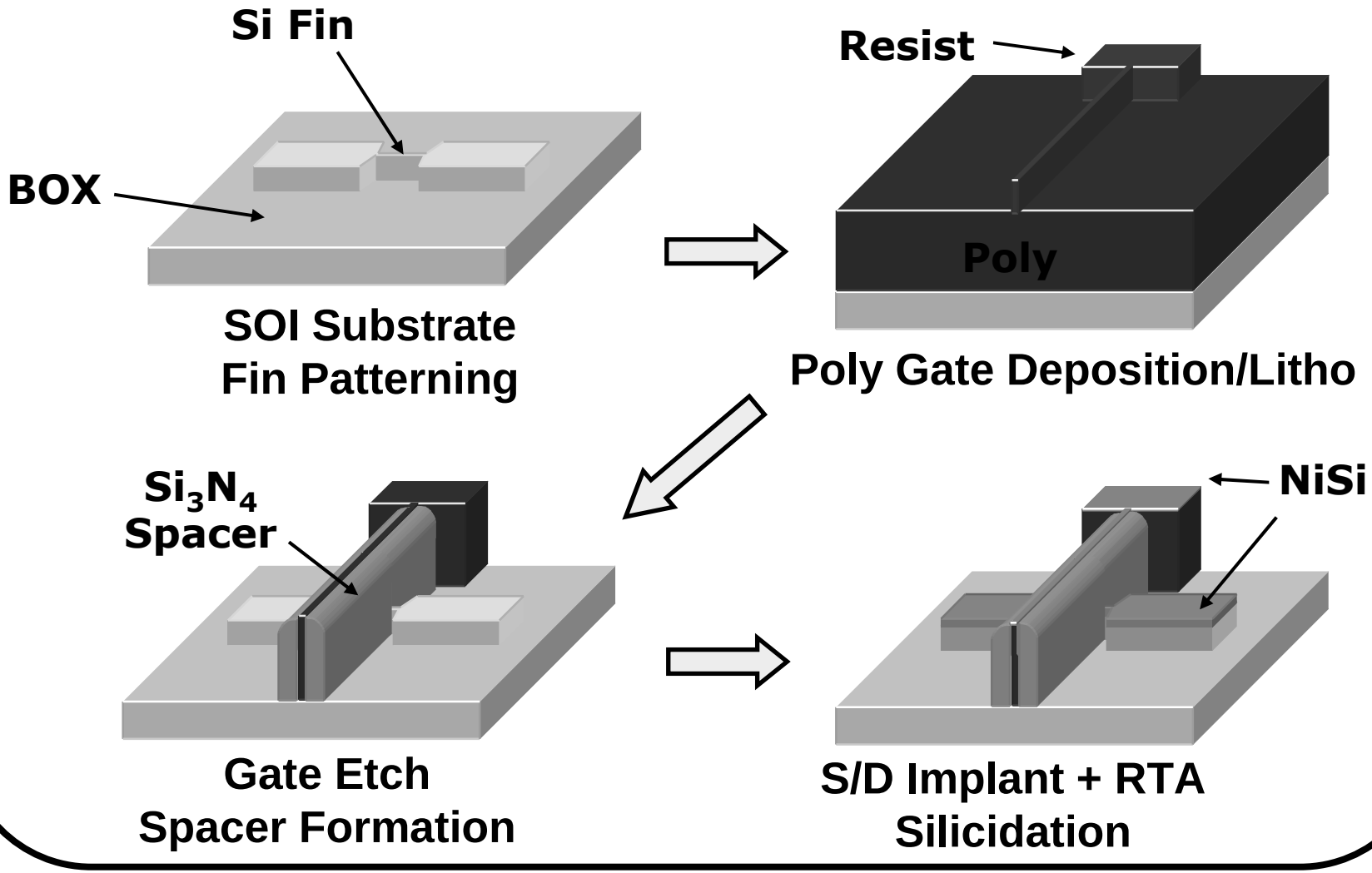
- One multi-gate structure, called **FinFET,** is particularly attractive for its simplicity of fabrication.
- Called FinFET because its silicon body resembles the back fin of a fish.
- The channel consists of the two vertical surfaces and the top surface of the fin.

Question: What is the channel width, W?

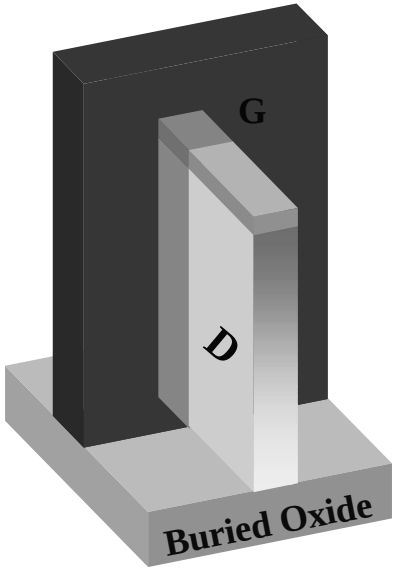
Answer: The sum of twice the fin height and the width of the fin.



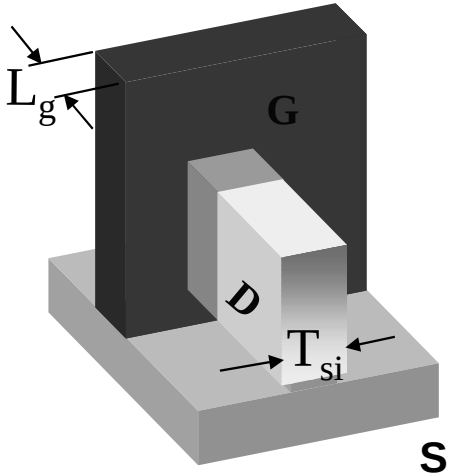
FinFET Process Flow



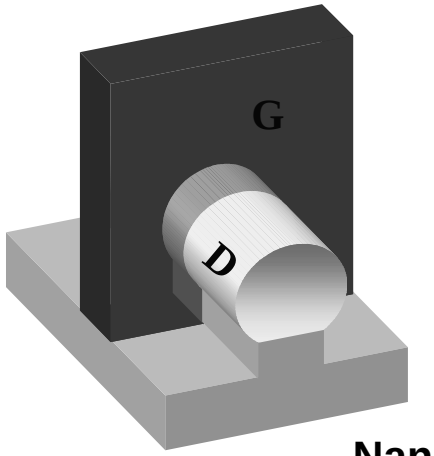
Variations of FinFET



Tall FinFET



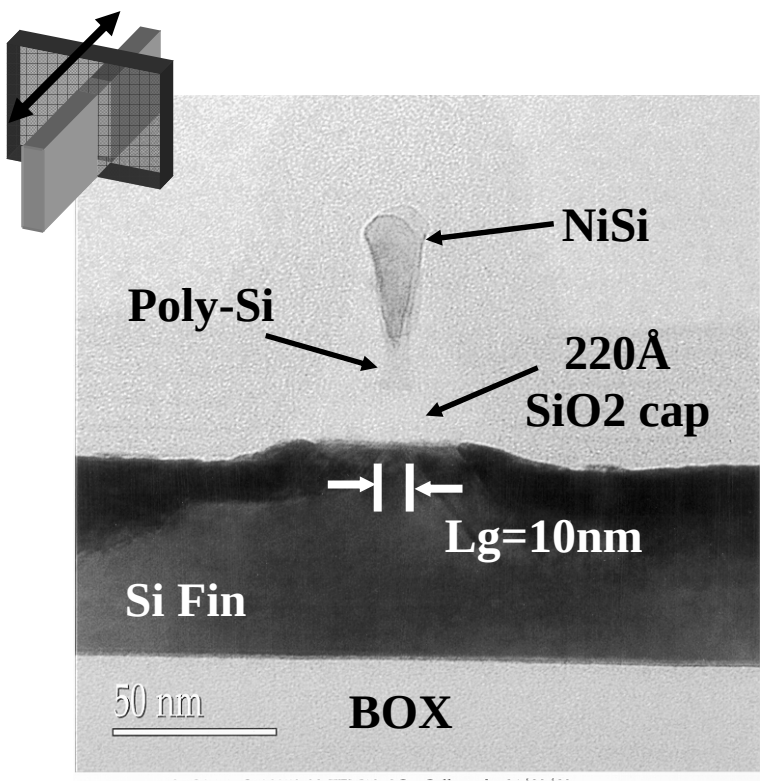
Short FinFET

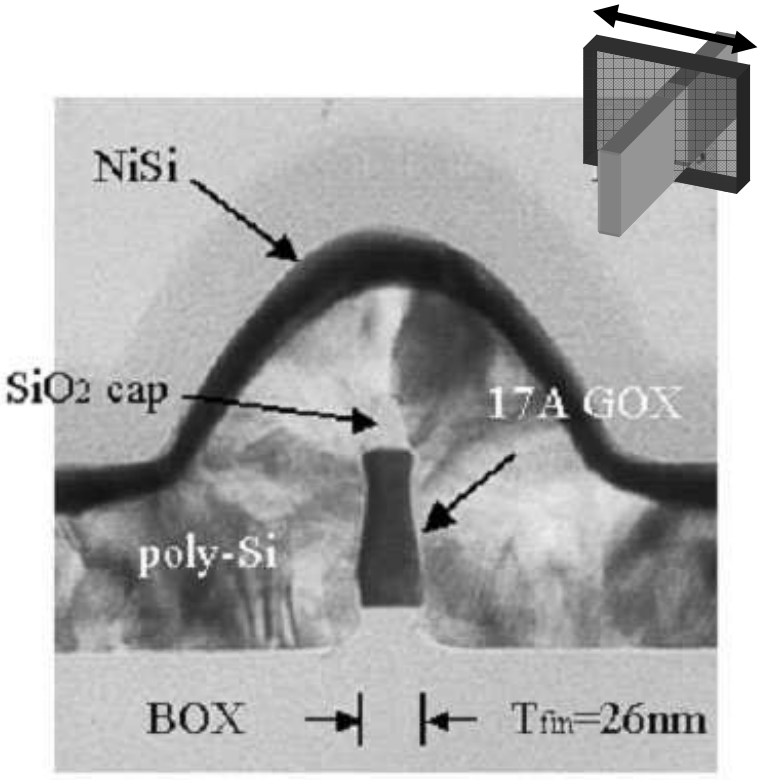


Nanowire FinFET

- **Tall FinFET** has the advantage of providing a large W and therefore large I_{on} while occupying a small footprint.
- **Short FinFET** has the advantage of less challenging lithography and etching.
- **Nanowire FinFET** gives the gate even more control over the silicon wire by surrounding it.

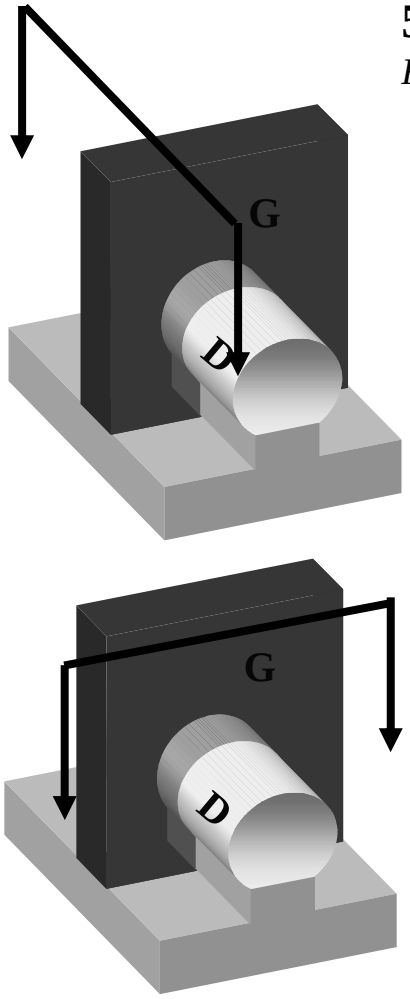
Tall FinFET with L_g=10nm





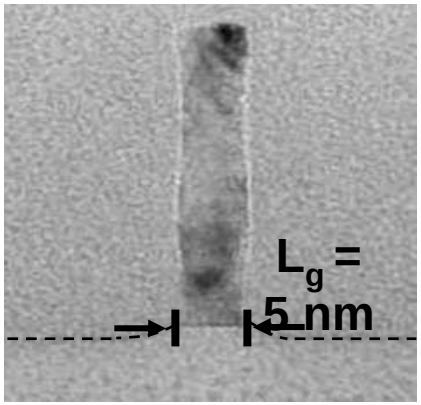
B. Yu *et al.*, *IEDM* 2002

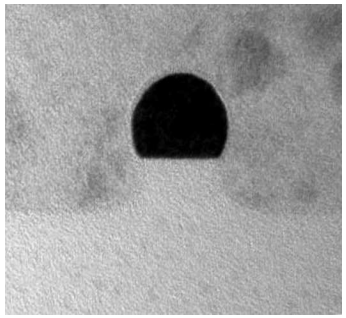
Nanowire FinFET



5nm Gate Length

F. Yang et al., 2004 VLSI Tech Symp.



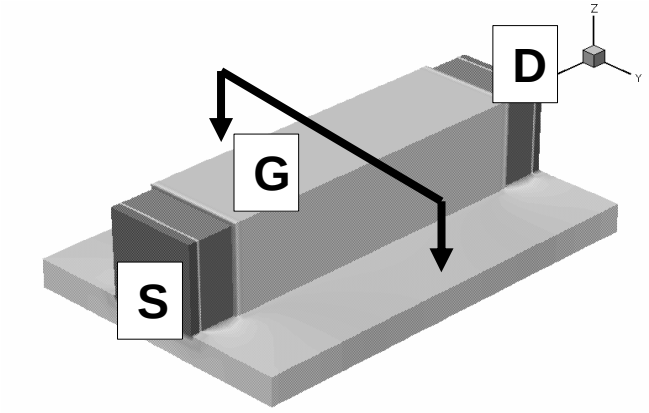


Device Simulation and Process Simulation

- **Device Simulation**
  - Commercially available computer simulation tools can solve all the equations presented in this book simultaneously with few or no approximations.
  - Device simulation provides quick feedback about device design before long and expensive fabrication.
- **Process Simulation**
  - Inputs to process simulation: lithography mask pattern, implantation dose and energy, temperatures and times for oxidization and annealing steps, etc.
  - The process simulator generates a 2-D or 3-D structures with all the deposited or grown and etched thin films and doped regions.
  - This output may be fed into a device simulator as input together with applied voltages.

Example of Device Simulation---

Density of Inversion Charge in the Cross-Section of a FinFET Body

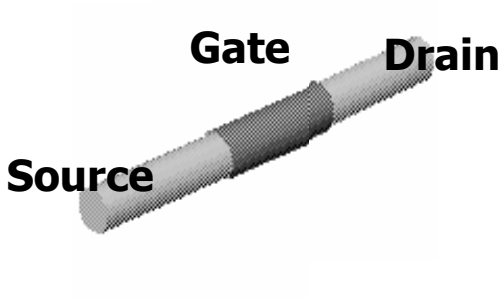


C.-H. Lin et al., 2005 SRC TECHCON

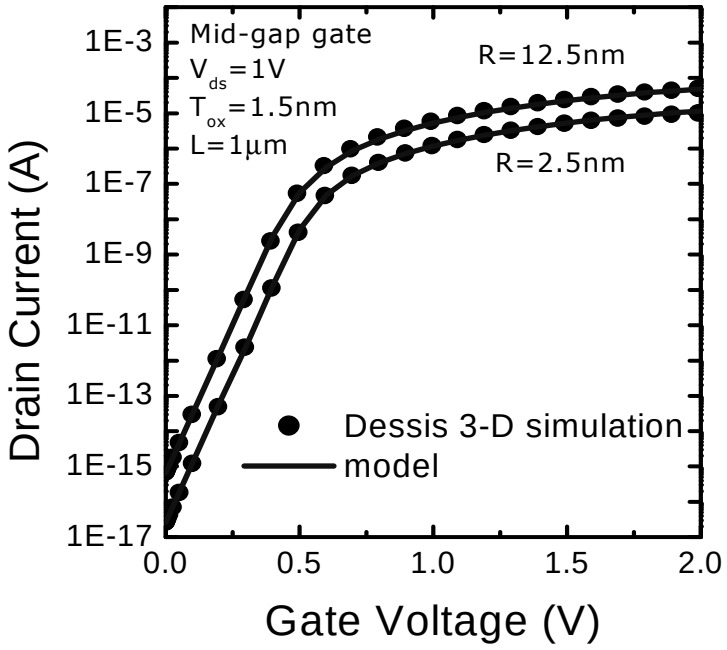
Tall FinFET **Short FinFET**

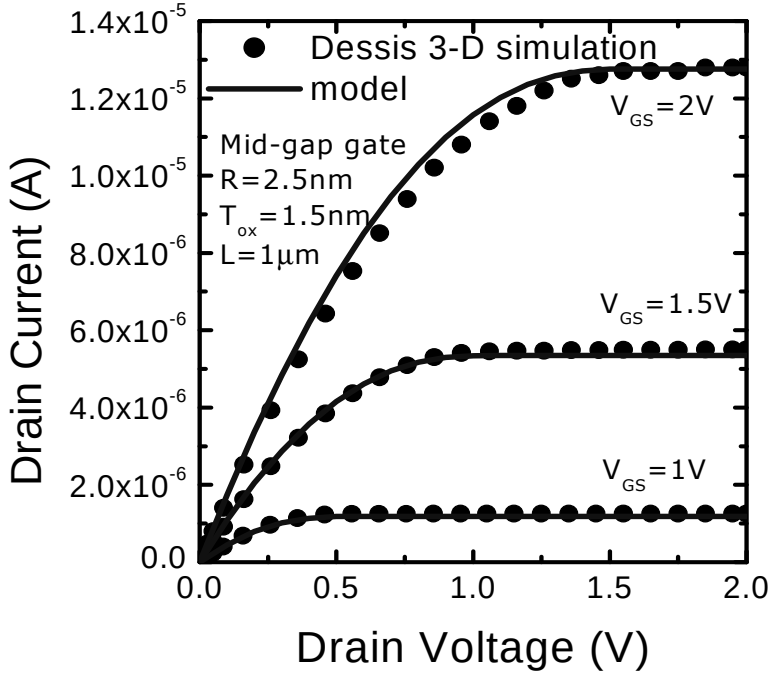
- The inversion layer has a significant thickness (T_{inv}).
- There are more more subthreshold inversion electrons at the corners.

I-V of a Nanowire "Multi-Gate" MOSFET



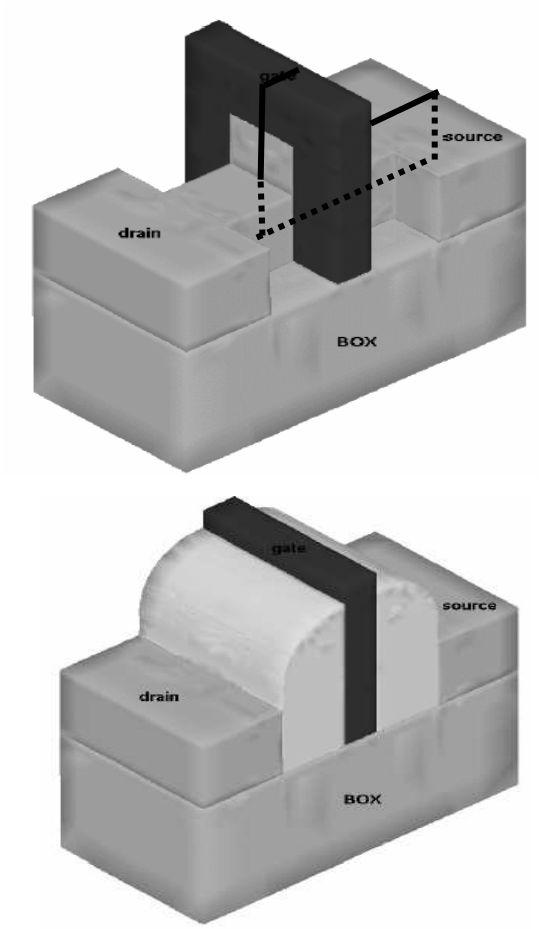
C.-H. Lin et al., 2005 SRC TECHCON

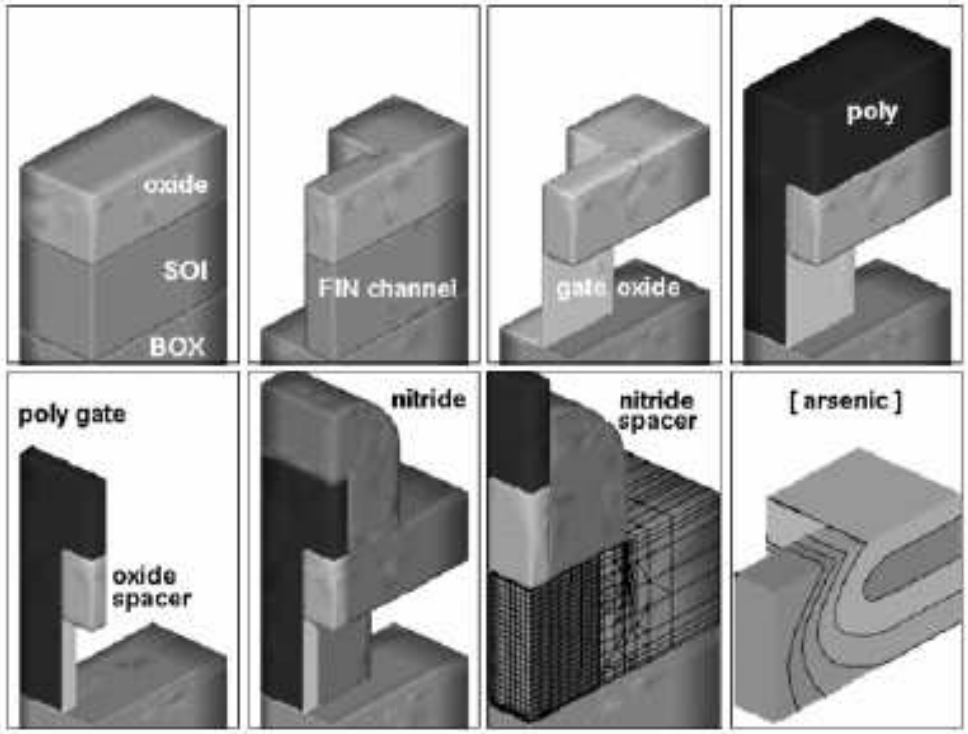




Example of Process Simulation

- **FinFET Process**





The small figures only show 1/4 of the complete FinFET--the quarter farthest from the viewer.

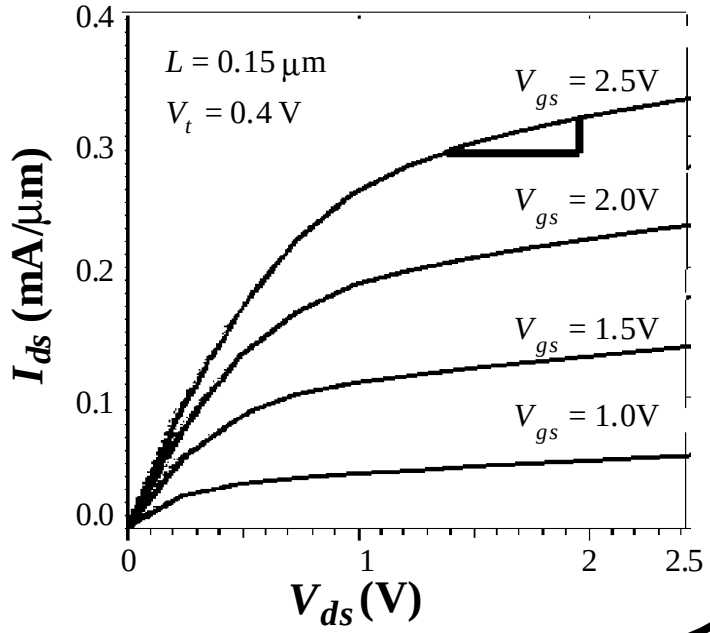
Manual, Taurus Process, Synoposys Inc.

7.9 Output Conductance

- I_{dsat} does NOT saturate in the saturation region, especially in short channel devices!
- The slope of the I_{ds}-V_{ds} curve in the saturation region is called the **output conductance (g_{ds})**,

$$g_{ds} \equiv \frac{dI_{dsat}}{dV_{ds}}$$

- A smaller g_{ds} is desirable for a large voltage gain, which is beneficial to analog and digital circuit applications.



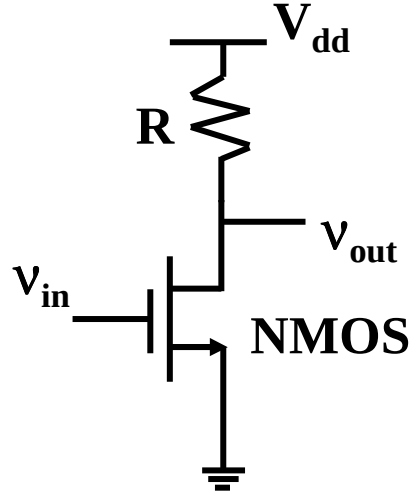
Example of an Amplifier

- The bias voltages are chosen such that the transistor operates in the saturation region. A ***small signal*** input, v_{in}, is applied.

$$i_{ds} = g_{msat} \cdot v_{gs} + g_{ds} \cdot v_{ds}$$
$$= g_{msat} \cdot v_{in} + g_{ds} \cdot v_{out}$$

$$i_{ds} = -v_{out} / R \cdot$$

$$\Rightarrow \quad v_{out} = \frac{-g_{msat}}{(g_{ds} + 1/R)} \times v_{in}$$



- The voltage gain is $g_{msat}/(g_{ds} + 1/R)$.
- A smaller g_{ds} is desirable for large voltage gain.
- Maximum available gain is g_{msat}/g_{ds}

What Parameters Determine the g_{ds} ?

$$g_{ds} \equiv \frac{dI_{dsat}}{dV_{ds}} = \frac{dI_{dsat}}{dV_t} \cdot \frac{dV_t}{dV_{ds}}$$

$$\frac{dI_{dsat}}{dV_t} = \frac{-dI_{dsat}}{dV_{gs}} = -g_{msat} \quad \textbf{and} \quad \frac{dV_{ds}}{dV_{ds}} = e^{-L/l_d}$$

I_{dsat} *is a function of* V_{gs}-V_t

(From Eq. 7.3.3, $V_t = V_{t-long} - V_{ds} \cdot e^{L/l_d}$ *)*

$$\Rightarrow \quad g_{ds} = g_{msat} \times e^{-L/l_d}$$

$$\Rightarrow \quad \text{Max voltage gain } (R \to \infty) = \frac{g_{msat}}{g_{ds}} = e^{L/l_d}$$

- A larger L or smaller l_d, i.e. smaller Tox, Wdep, Xj, can increase the maximum voltage gain.
- The cause is "V_t dependence on V_{ds}"in short channel transistors.

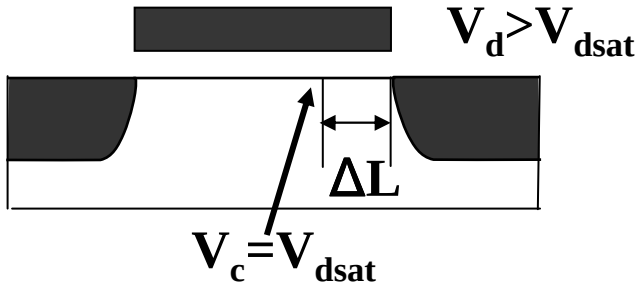
Channel Length Modulation

• For large L and V_{ds} close to V_{dsat}, another mechanism may be the dominant contributor to g_{ds}. That is **channel length modulation**.

• A voltage drop, V_{ds}-V_{dsat}, is dissipated over a finite distance next to drain, causing the "channel length"to decrease. More with increasing V_{ds}.

$$g_{ds} = \frac{I_d \cdot I_{dsat}}{L(V_{ds} - V_{dsat})}$$

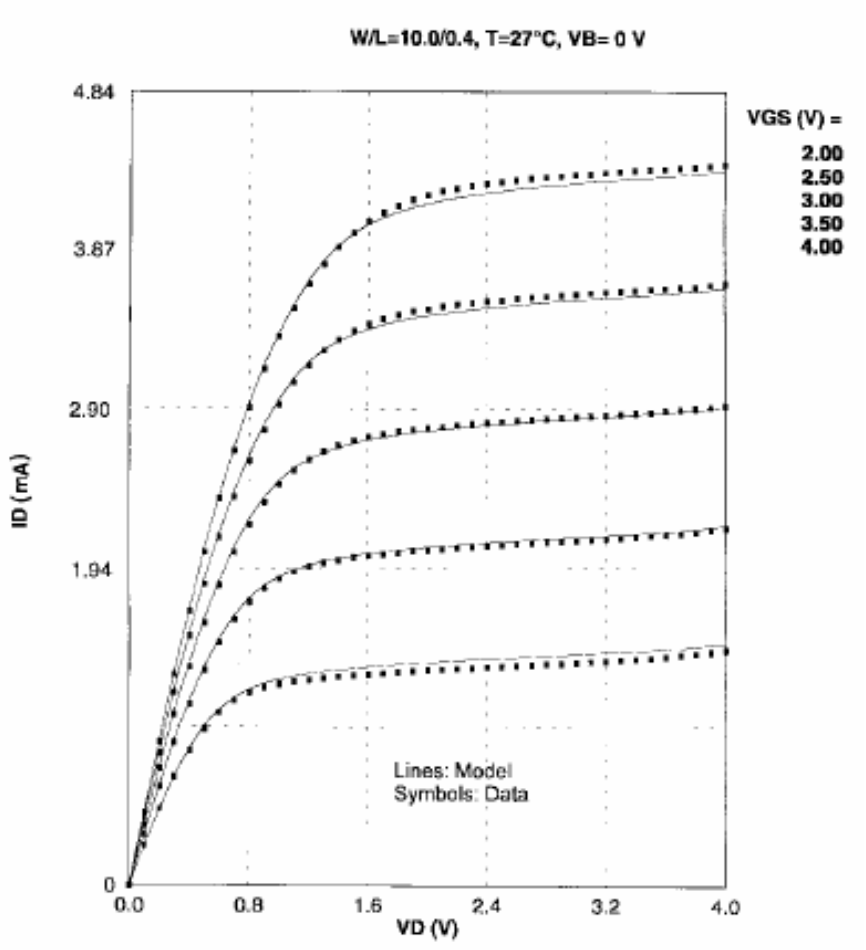
$$I_d \approx \sqrt[3]{T_{ox} W_{dep} X_j}$$

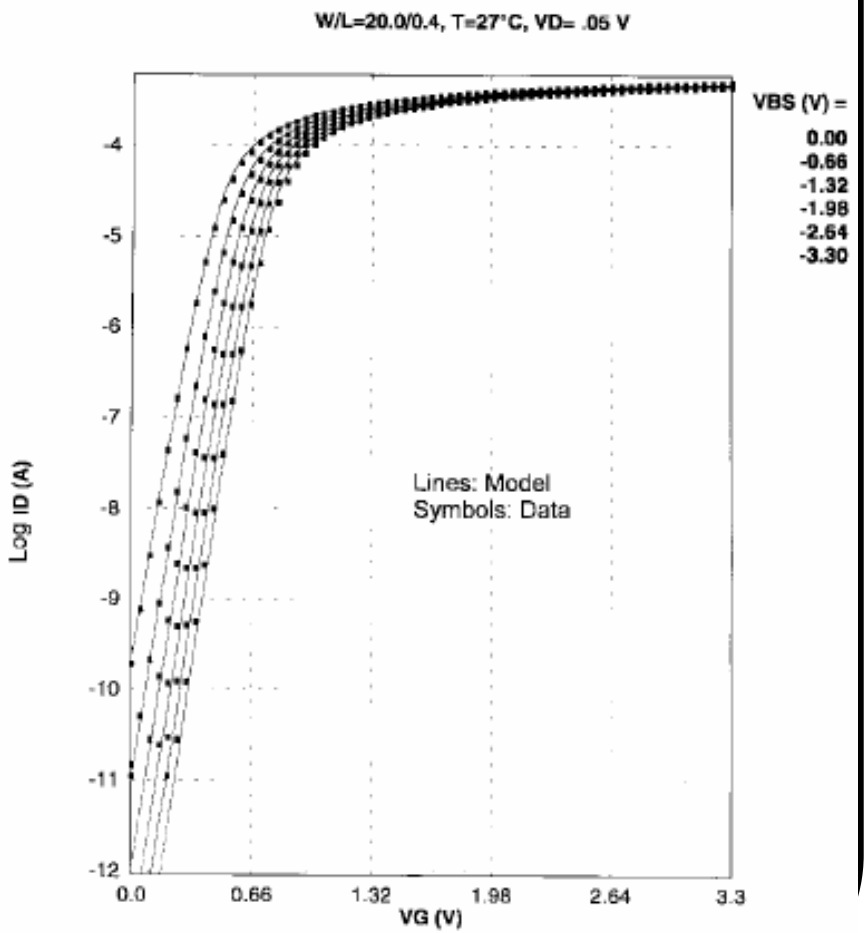


7.10 MOSFET Modeling for Circuit Simulation

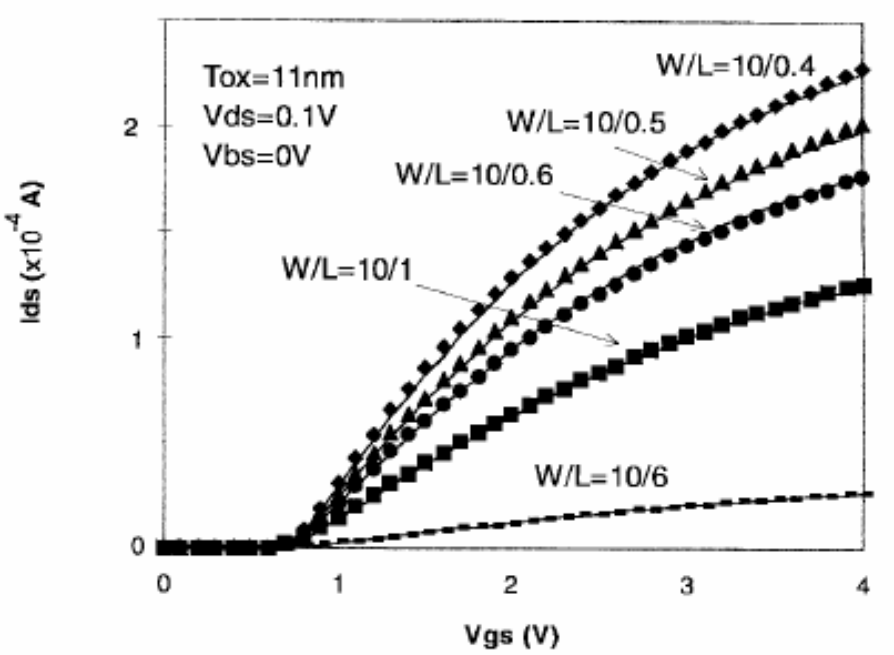
- For circuit simulation, MOSFETs are modeled with analytical equations.
- Device model is the link between technology/manufacturing and design/product. The other link is design rules.
- Circuits are designed A. through circuit simulations or B. using cell libraries that have been carefully designed beforehand using circuit simulations.
- **BSIM** (**B**erkeley **S**hort-channel **I**GFET **M**odel) is the world's first industry standard MOSFET model. It contains all the equations presented in these chapters.

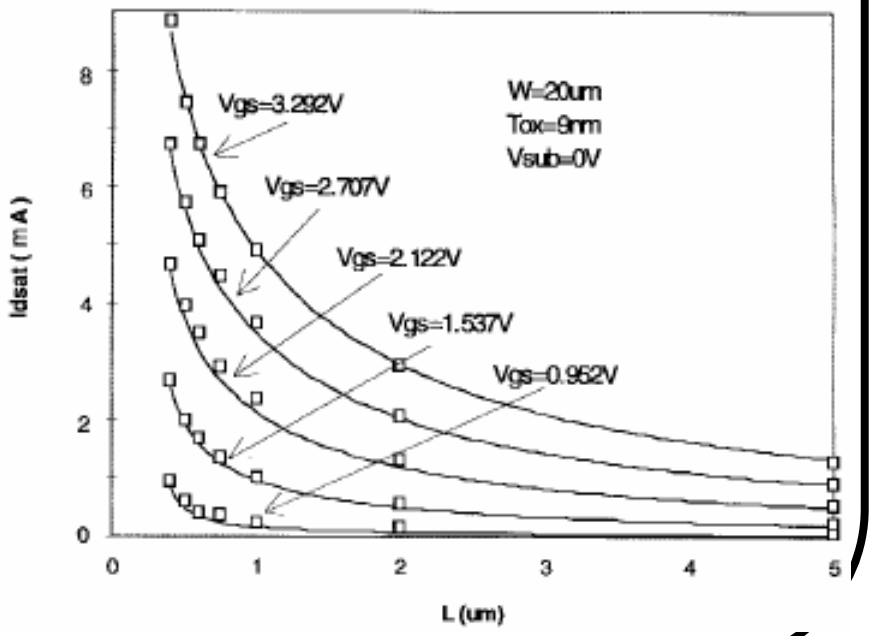
Examples of BSIM Model Results





Example of BSIM Model Results





Example of BSIM Model Results