MOS Transistor (Scaling)

A) Overview
B) Short channel effects (use TCAD to explore)
C) Idealistic (constant field) and realistic scaling
D) Innovation in material and architecture for scaling
E) Variability
F) Future challenges

Technology Scaling

Scale transistor size → Faster switching, more transistors/area

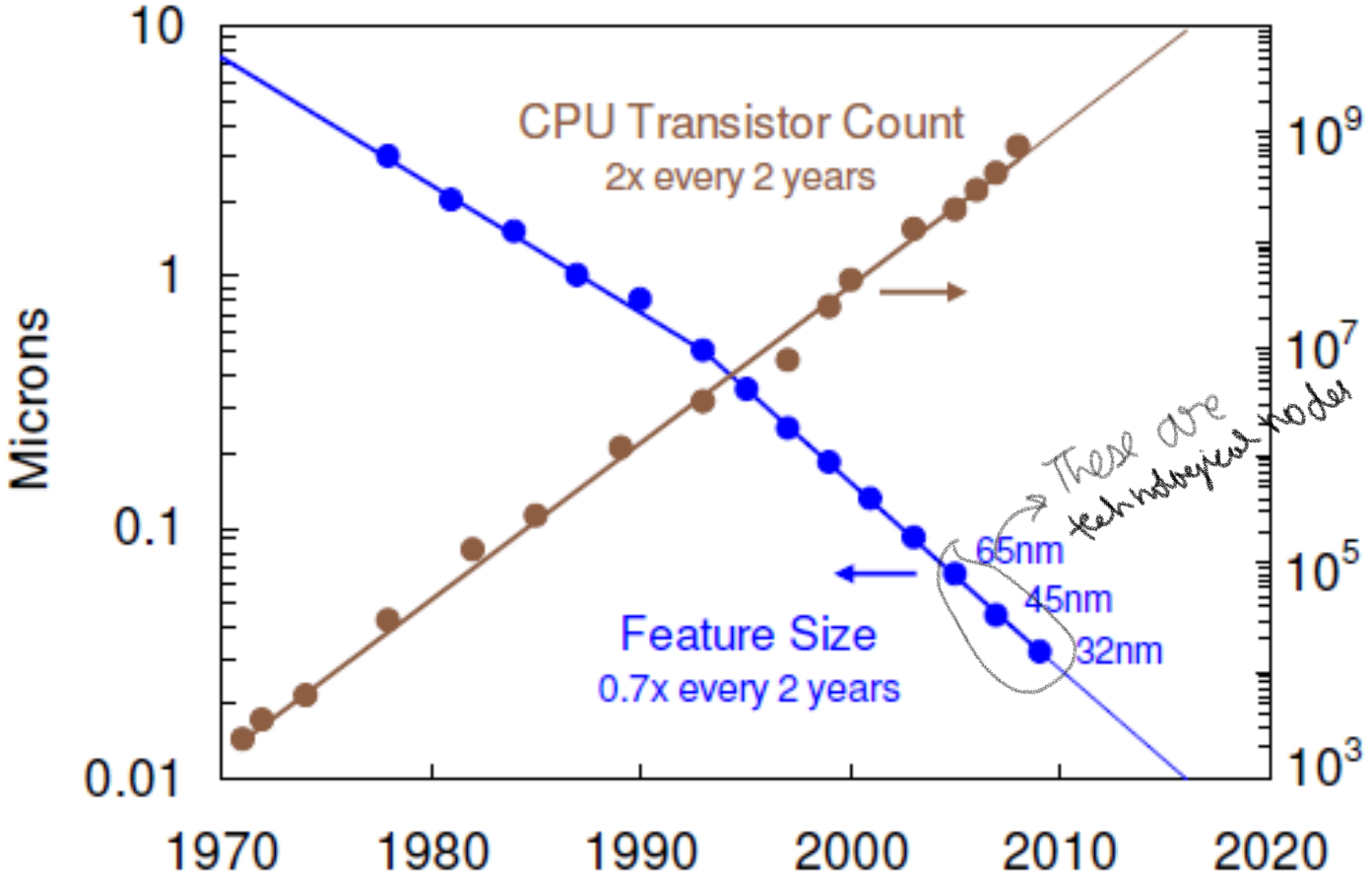
Higher performance, more functionality, lower supply voltage (mobile applications)

Today's mobile phones can do much more than yesterday's PCs

Huge reduction in cost / functionality

Transistor Scaling

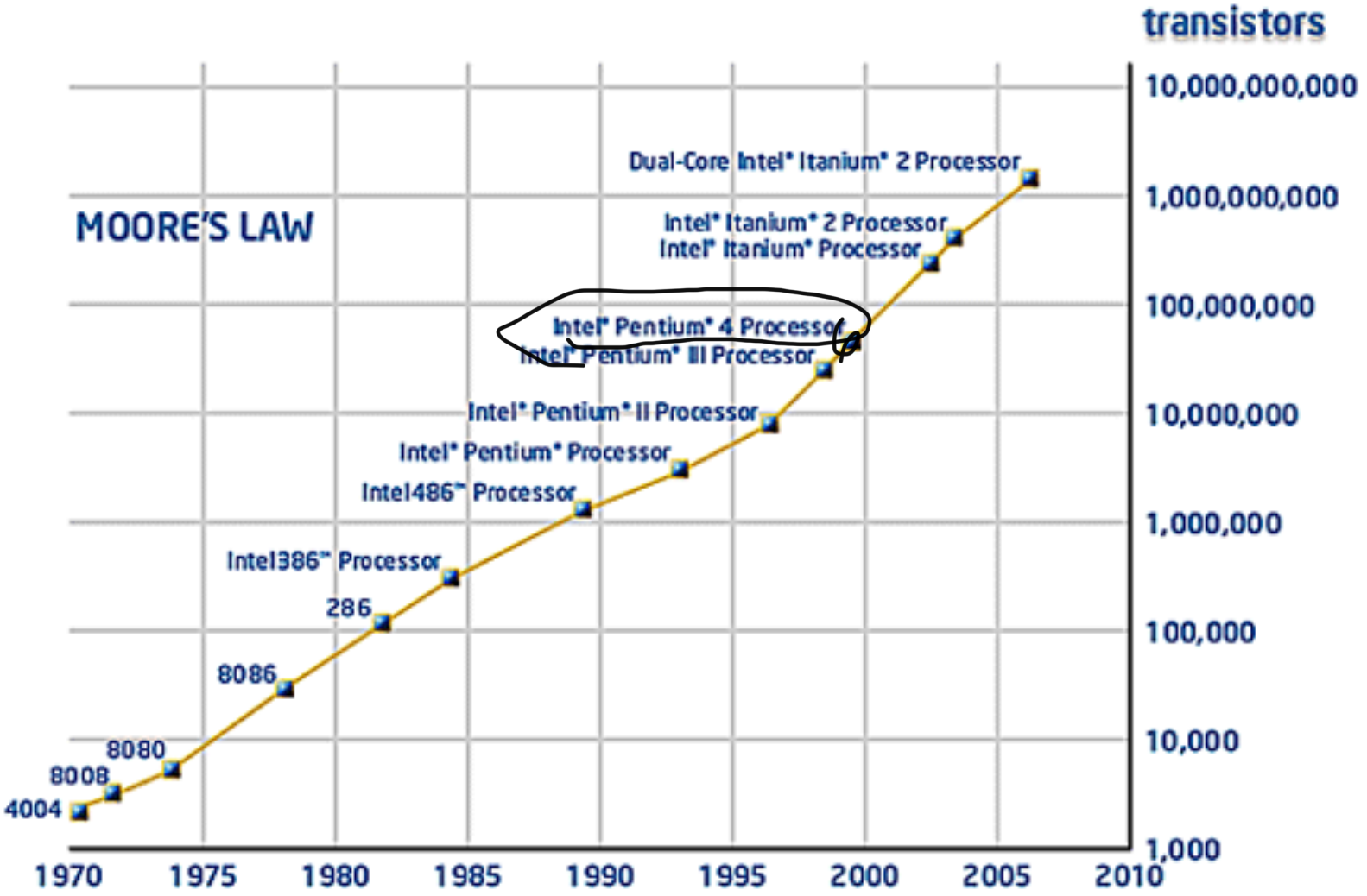
Exponential increase over the last ~40 years



Intel

Transistor Count (Functionality)

Exponential increase over the last ~40 years



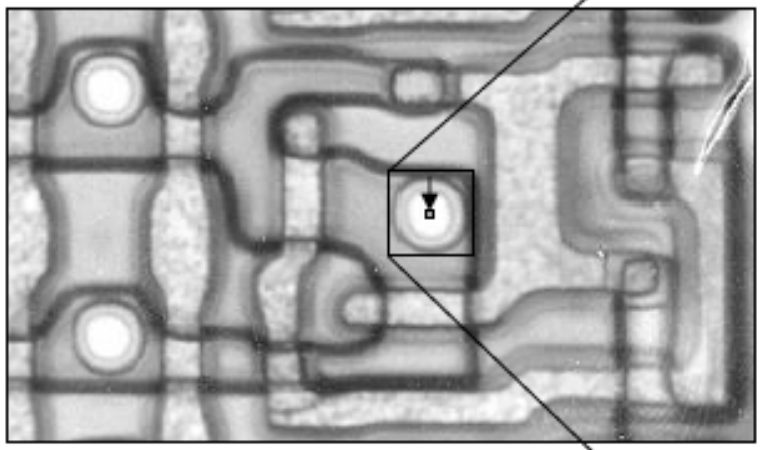
Intel

Example of Scaling – 6T SRAM

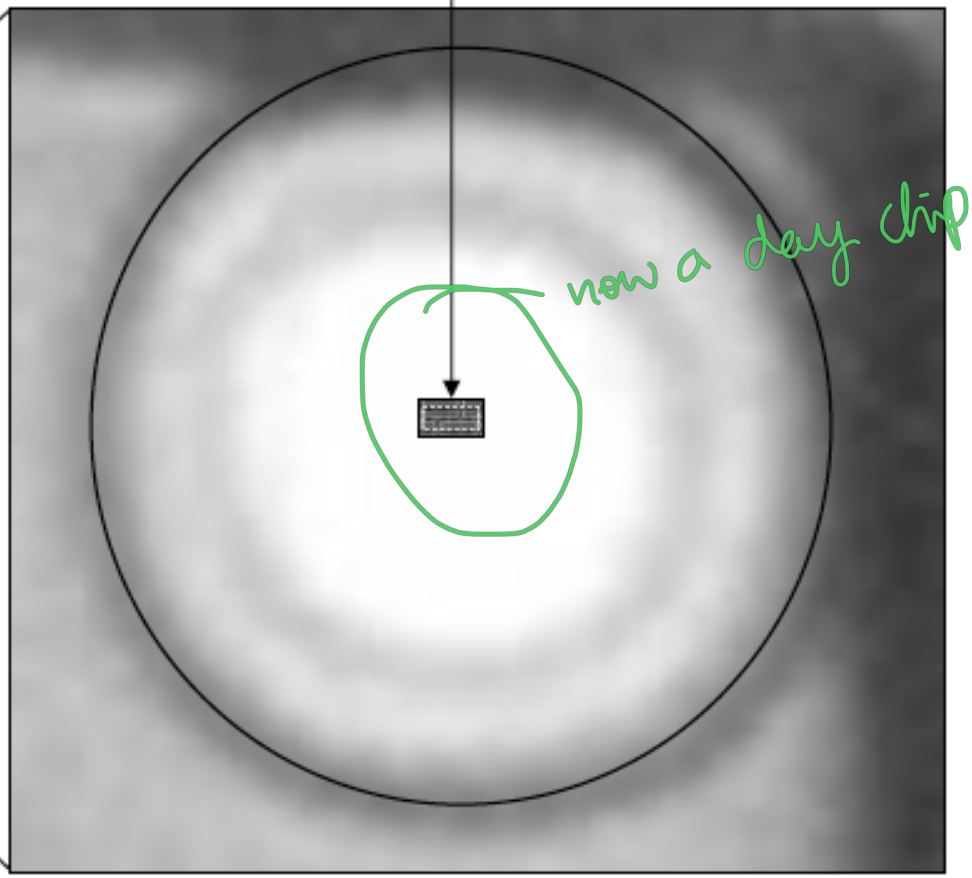
1980 SRAM Cell: 1700 um²

22nm SRAM Cell: 0.092 um²

Intel



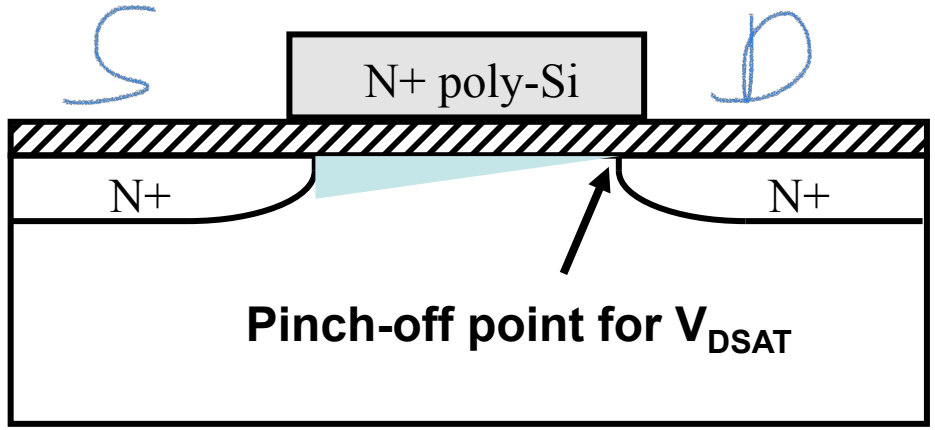
1 10000X

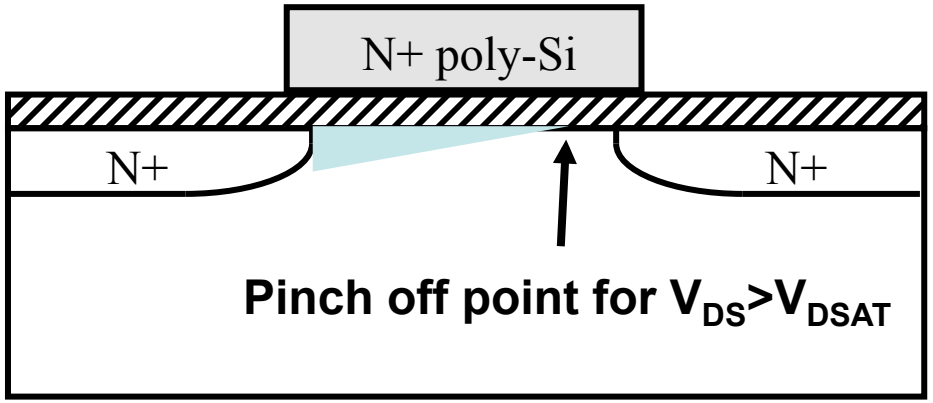


Small enough that a 2011 22nm SRAM cell is dwarfed by a 1980 SRAM cell CONTACT

Channel Length Modulation

Beyond pinch-off ($V_{DS}>V_{DSAT}$), pinch off point moves towards S



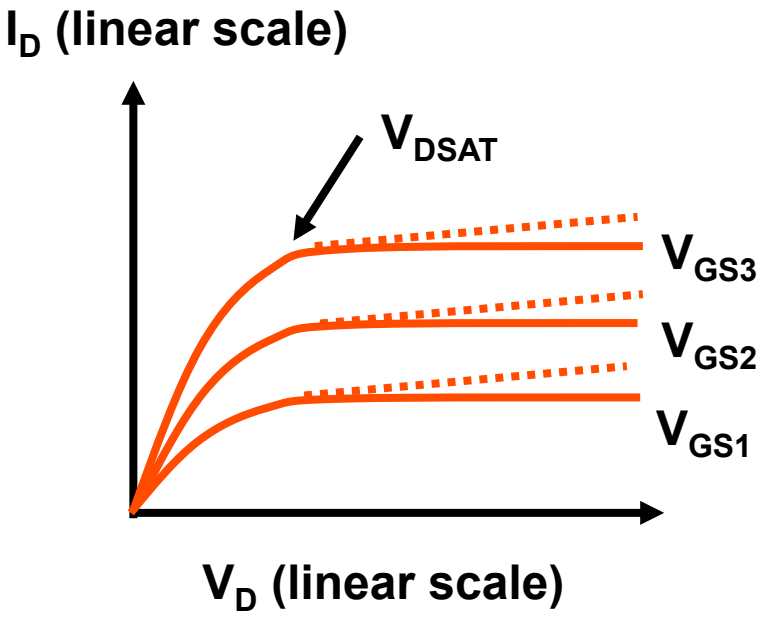


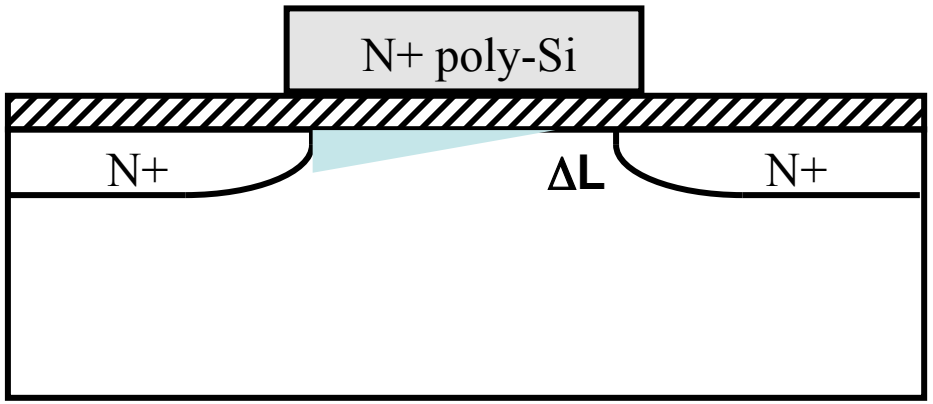
Channel length get modulated (shortened) by ΔL (pure depletion layer), voltage drop in ΔL is (V_{DS}-V_{DSAT})

Current conduction similar to reverse-bias PN junction

Channel Length Modulation

Reduction in L_{eff} to $L_{eff} - \Delta L$





$$\Delta L = \sqrt{\frac{2\varepsilon_{si}(V_D - V_{DSAT})}{qN_A}}$$

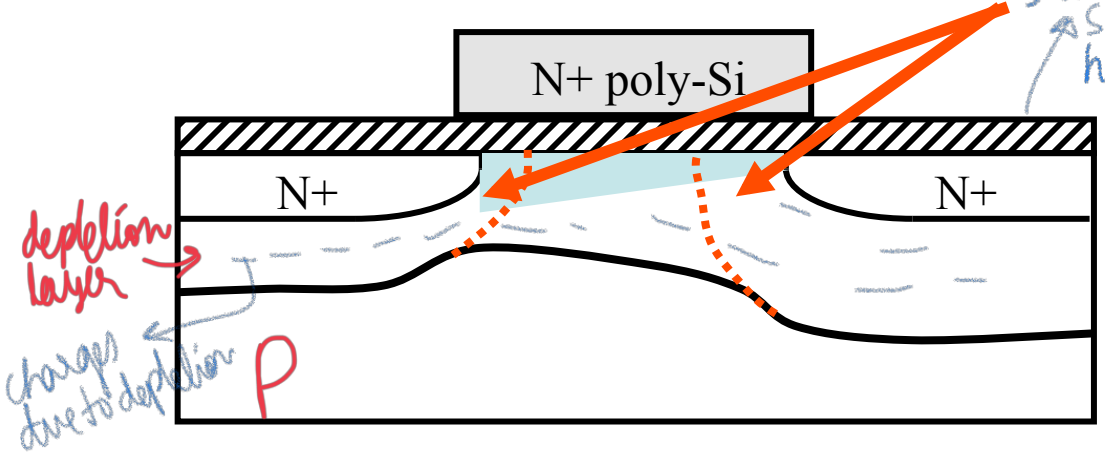
change in channel length due to pinch off

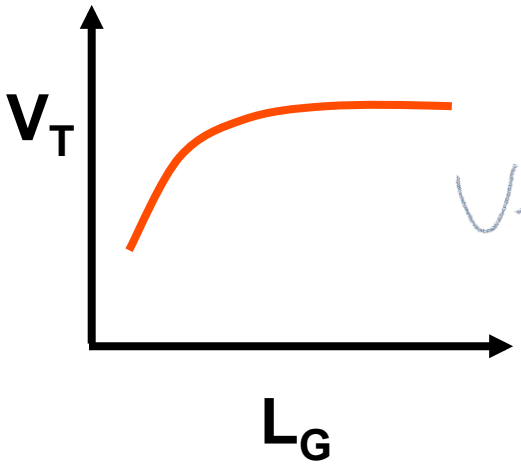
$$I_D = \mu_{eff} C_{OX} \frac{W}{L_{eff} - \Delta L} \frac{(V_{GS} - V_T)^2}{2m}$$

Slight increase in I_D for $V_{DS} > V_{DSAT}$

Charge Sharing by S / D (V_T Roll-Off)

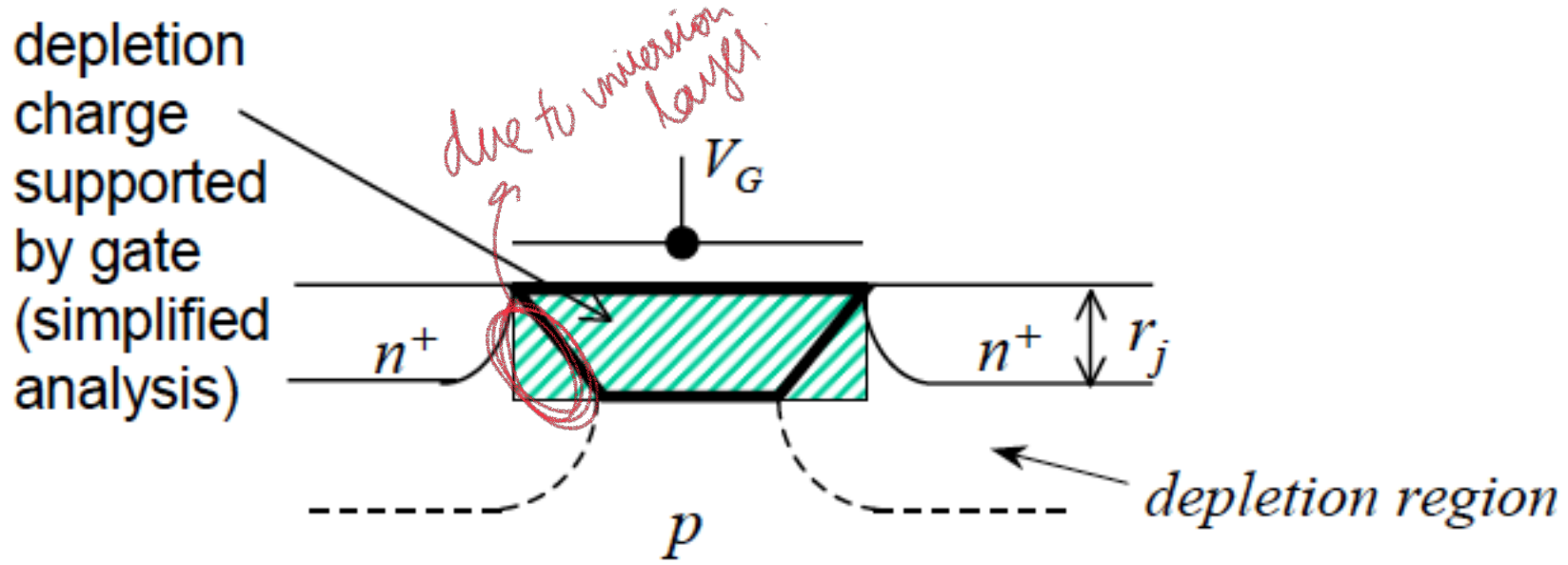
Charge in channel near S & D are supported by gate and also by depletion region inside S & D





Due to charge sharing by S & D, charges need to be supported by gate is less → V_{GS} required to achieve V_T is less → Significant effect as L_G is reduced → V_T roll off

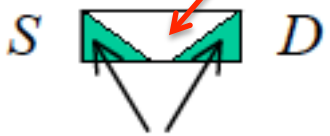
Charge Sharing by S / D (V_T Roll-Off)



Large L:

Depletion charge supported by S/D

Small L:



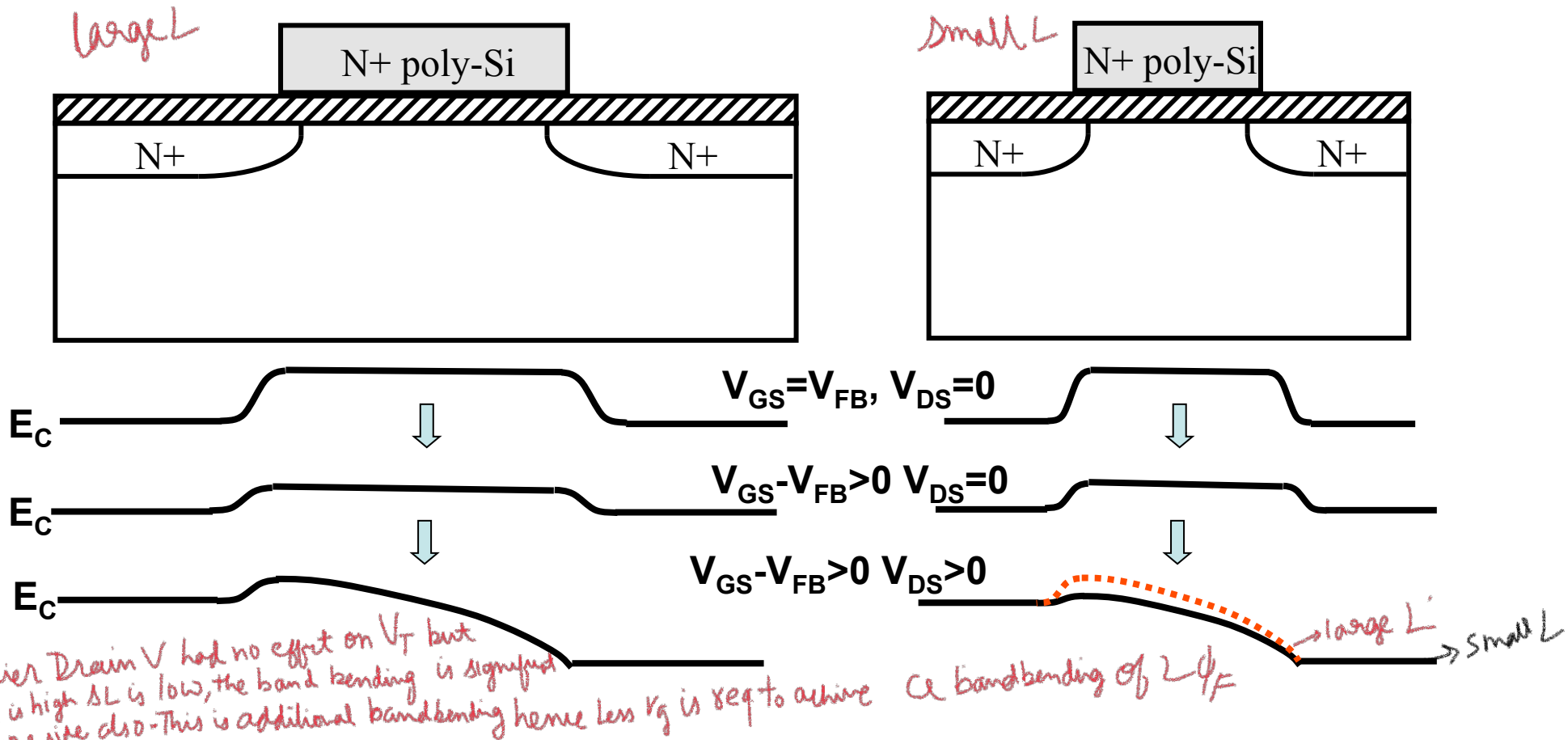
Lower charge supported by gate

Depletion charge supported by S/D

cannot neglect charge of depletion layer

Totalcharge = Q by gate + Q by S/D
≃ Q by gate (long L)

Drain Induced Barrier Lowering (DIBL)



Earlier Drain V had no effect on V_T but when V_{DS} is high & L is low, the band bending is significant on the source side also - This is additional bandbending hence less V_G is req to achieve ie bandbending of $2\phi_F$

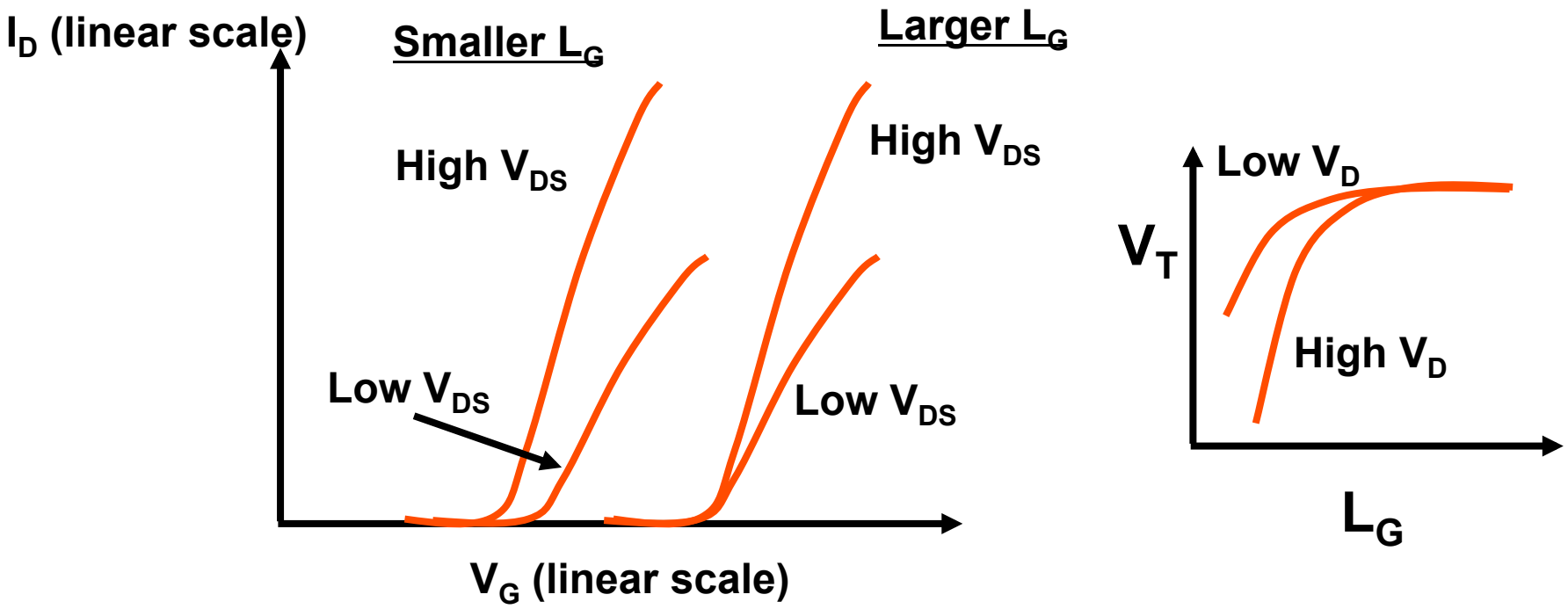
For small L_G, V_{DS} causes additional barrier lowering at the S end → Reduced amount of barrier to lowered by V_{GS} to achieve V_T → V_T (saturation) is lower than V_T (linear)

Why

V_T Roll-Off and DIBL

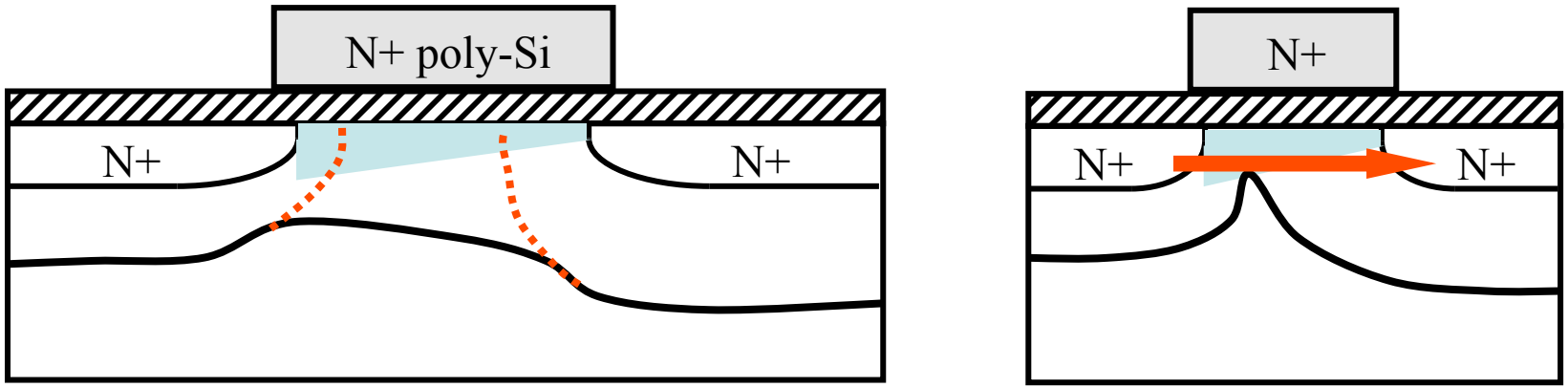
V_T roll-off → reduction in V_T at smaller L_G

DIBL → reduction in V_T at higher VDS (especially for smaller L_G)



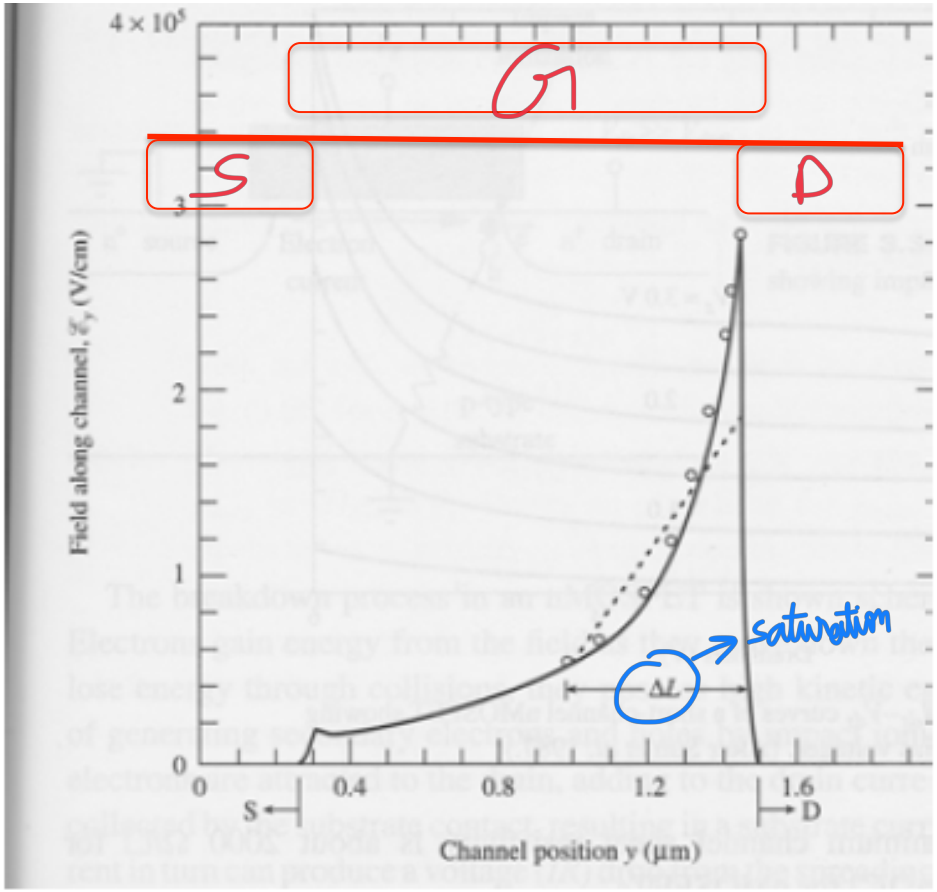
Punch Through

For short L_G at high V_{DS}, S and D depletion layer can touch each other (away from interface) → direct flow of carriers from S to D with no control of G



Punch through is a sub-surface phenomenon (happens away from the interface) → Same impact on V_T as DIBL

Lateral Electric Field



Dimension scaling > Voltage Scaling (discussed later)

High lateral electric field near the drain junction

(1) Velocity saturation

(2) Impact ionization

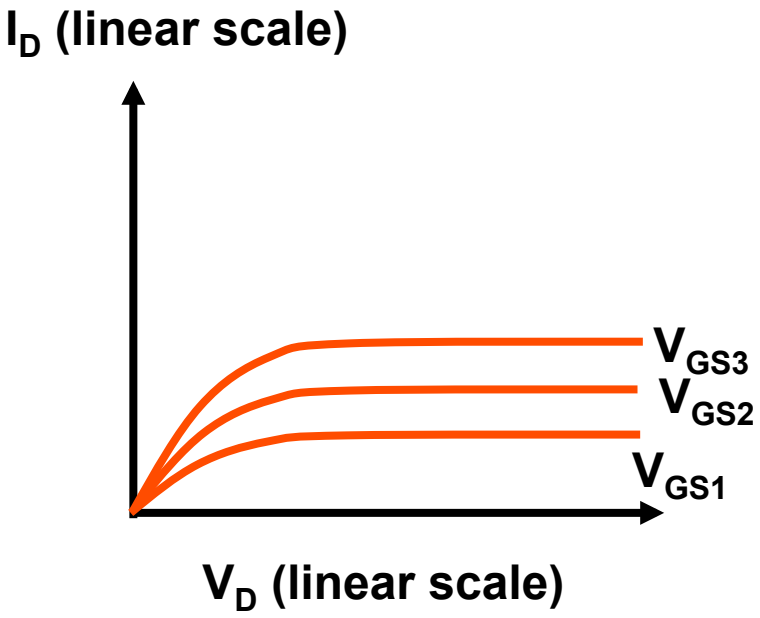
(3) Hot Carrier Injection

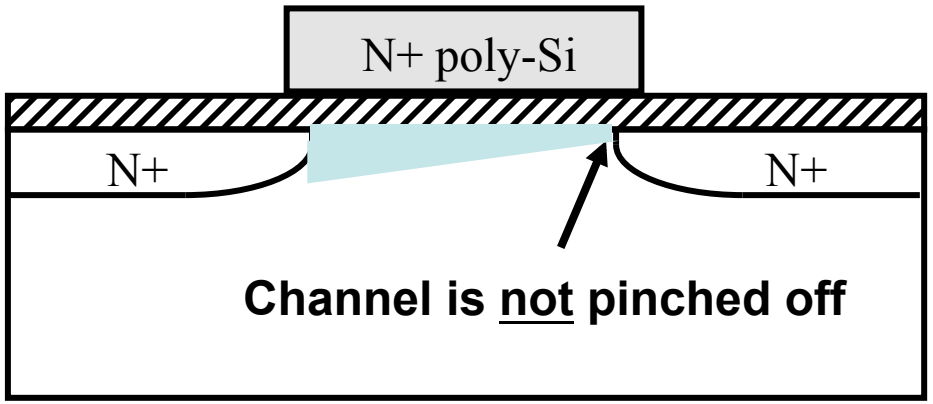
(4) Parasitic bipolar action

Ref: Taur and Ning

Velocity Saturation

For short L_G, E_y close to D junction becomes very high as V_{DS} is increased





Electron velocity saturates (does not increase with E_y)

$$I_{DSAT} \sim Q v_{sat} \sim C_{OX}(V_{GS} - V_T) v_{sat}$$

v_{sat}: saturation velocity

I_{DSAT} varies linearly with overdrive (not square law) $V_{GS} - V_T$

Mobility at High Vertical and Lateral Field

Reduction in mobility due to transverse field

$$\mu_{eff} = \frac{\mu_0}{1 + \alpha(V_{GS} - V_T)}$$

Reduction in mobility due to lateral field (reduction in carrier velocity)

$$v = \frac{\mu_{eff} E}{[1 + (E / E_c)^n]^{1/n}}$$

n=2 for electrons
n=1 for holes
vsat = μ_{eff} E_c

Analytical Solution for n=1

$$I_D = -WQ_i(V)\frac{\mu_{eff}\, dV/dy}{[1+(\mu_{eff}/v_{sat})]dV/dy}$$

$$I_D = -[\mu_{eff} WQ_i(V) + \frac{\mu_{eff} I_D}{v_{sat}}]\frac{dV}{dy}$$

$$I_D = \frac{-\mu_{eff}(W/L)\int_0^{V_{DS}} Q_i(V)\,dV}{1+(\mu_{eff} V_{DS}/v_{sat}L)}$$

Integrate from 0 to L and 0 to V_{DS}

$$Q_i(V) = -C_{OX}(V_G - V_T - mV)$$

Analytical Solution for n=1

$$I_D = \frac{\mu_{eff} C_{OX} (W/L)[V_G - V_T - \frac{m}{2} V_{DS}]V_{DS}}{1 + (\mu_{eff} V_{DS} / v_{sat} L)}$$

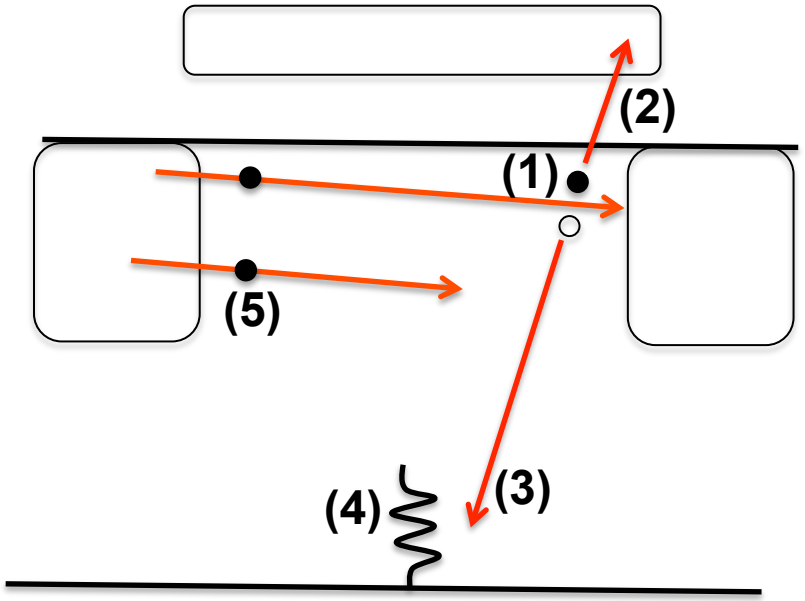
$$V_{DSAT} = \frac{2(V_G - V_T)/m}{1 + \sqrt{1 + 2\mu_{eff}(V_G - V_T)/(m v_{sat} L)}}$$

Solve for $dI_D/dV_{DS}=0$

$$I_{DSAT} = C_{OX} W v_{sat} (V_G - V_T) * \frac{\sqrt{1 + 2\mu_{eff}(V_G - V_T)/(m v_{sat} L)} - 1}{\sqrt{1 + 2\mu_{eff}(V_G - V_T)/(m v_{sat} L)} + 1}$$

Substitute V_{DSAT}

High Lateral Field Effects



(1) Impact ionization, creation of electrons and holes

(2) Electron injection into gate – gate current, oxide damage

(3) Hole injection towards substrate – substrate current

(4) Forward bias of source substrate junction

(5) Additional electron injection from source

Scaling → Necessities and Challenges

Reduce L_G → Higher speed, density

S & D junctions come closer → short channel effects (channel length modulation, velocity saturation, DIBL, V_T roll off, punch through)

Solution:

Scale vertical dimensions → T_{OX} & X_j

Increase doping, reduce operating voltage

Scaling → Necessities and Challenges

On current → keep it high

$$I_{DSAT} = \mu_{eff} C_{OX} \frac{W}{L_{eff}} \frac{(V_{GS} - V_T)^2}{2m}$$ **Classical**

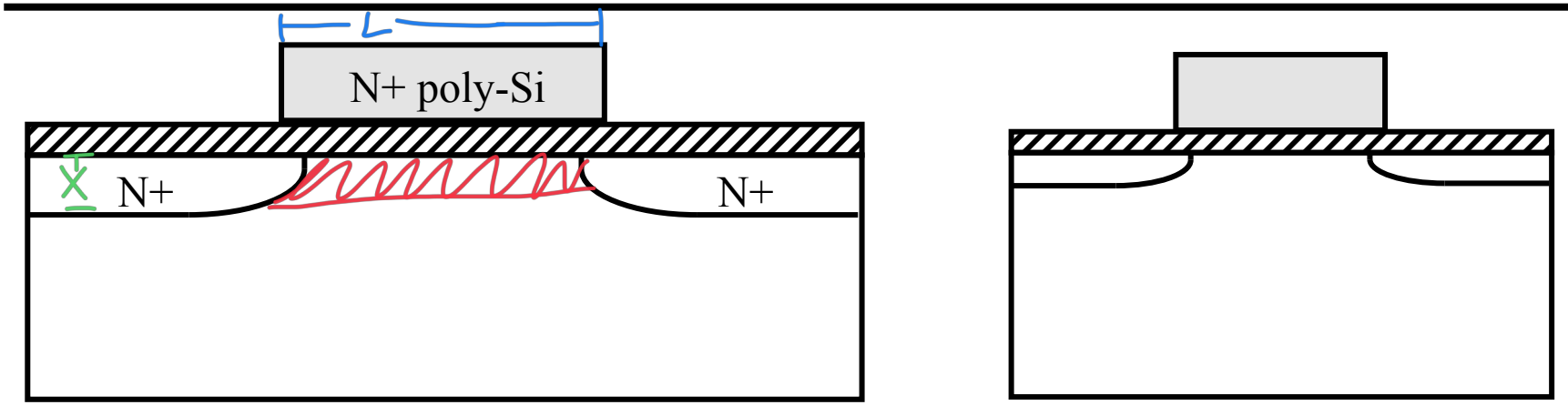
$$I_{DSAT} = \mu_{eff} C_{OX} v_{sat} (V_{GS} - V_T)$$ **Velocity saturation**

Off current → keep it low

$$I_D = \mu_{eff} C_{OX} \frac{W}{L_{eff}} (m-1)(\frac{kT}{q})^2 \, \mathrm{e}^{q(V_G - V_T)/mkT} [1 - e^{-qV_{DS}/kT}]$$

$$m = 1 + C_D / C_{OX} = 1 + 3T_{OX} / W_{DMAX}$$

Constant Electric Field Scaling Rules



Long channel		Short channel
W, L_G, T_{OX}, X_j	Dimensions	W/k, L_G/k, T_{OX}/k, X_j/k
N_A, N_D	Doping	N_A.k, N_D.k
V_{DD}	Supply voltage	V_{DD}/k

JUNCTION DEPTH

Dennard, IBM

Scaling of Device Parameters (k>1)

Electric field → 1 CONSTANT ELECTRIC FIELD SCALING

Carrier velocity → 1

Depletion layer width → 1/k

Capacitance → 1/k

Inversion charge density → 1

Drift current (on-state, per W) → 1

Diffusion current (off-state, per W) → k

Scaling of Circuit Parameters

Delay time (C.V/I) → 1/k

Power dissipation (V.I) → $1/k^2$

Power-delay product → $1/k^3$

Circuit density (1/Area) → k^2

Power density (P/A) → 1

Trouble: Non-Scaling Factors

Threshold voltage

$$V_T = V_{FB} + 2\phi_F + \frac{\sqrt{2q\varepsilon_{si}N_A(2\phi_F)}}{C_{OX}}; \qquad V_{FB} = -(E_G/2q + \phi_F)$$

Subthreshold current

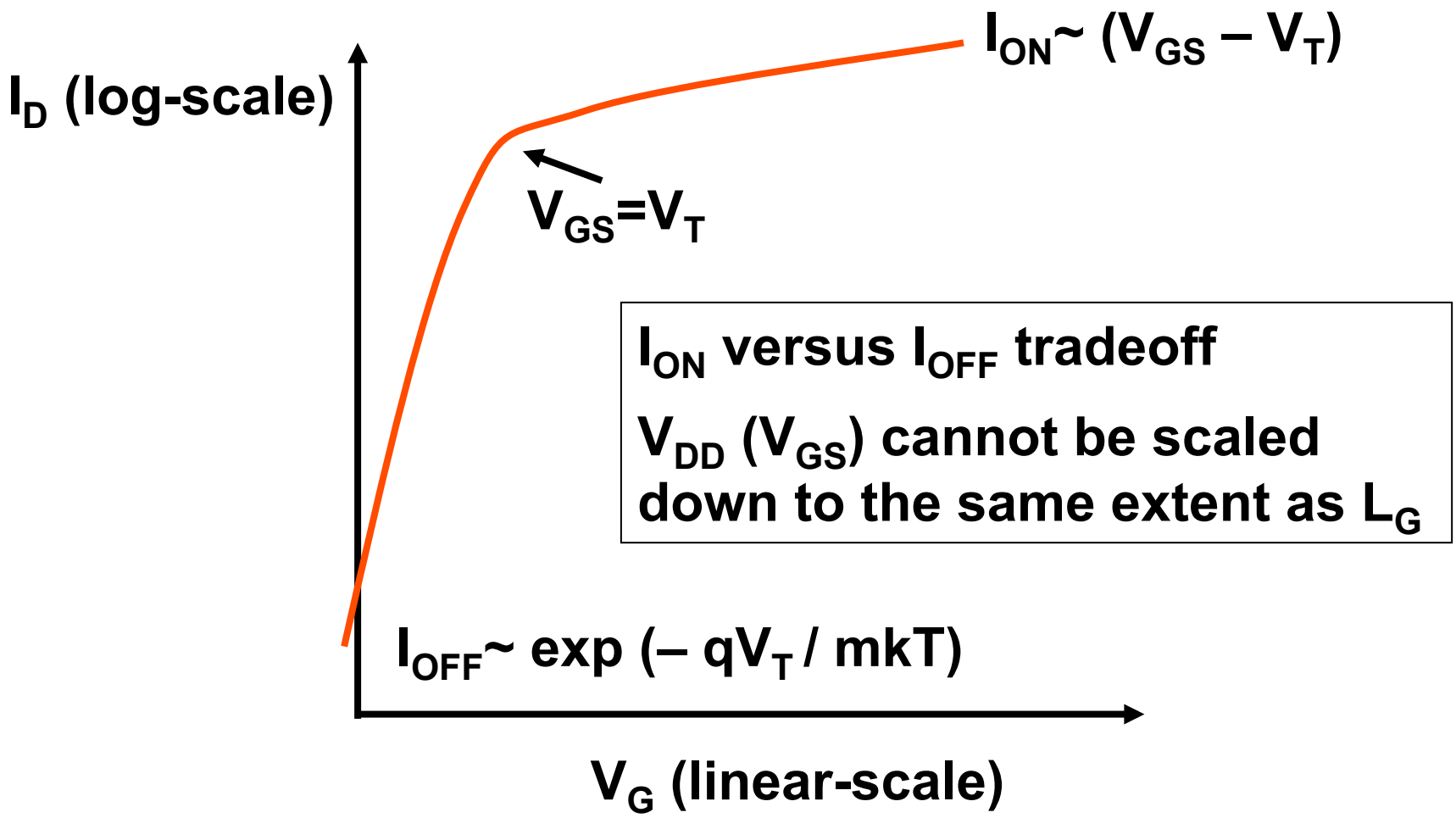
$$I_D = \mu_{eff}.C_{OX}.\frac{W}{L_{eff}}(m-1)\left(\frac{kT}{q}\right)^2 e^{q(V_{GS}-V_T)/mkT}(1-e^{-qV_{DS}/kT})$$

$$m = 1 + \frac{\sqrt{\varepsilon_{si}qN_A/4\phi_F}}{C_{OX}}$$

E_G, kT/q does not scale

The V_T scaling dilemma

Choice of proper V_T



Generalized scaling rules (k>1, α>1)

Long channel		Short channel
W, L_G, T_{OX}, X_j	Dimensions	W/k, L_G/k, T_{OX}/k, X_j/k
N_A, N_D	Doping	$N_A.(\alpha k)$, $N_D.(\alpha k)$
V_{DD}	Supply voltage	$V_{DD}.(\alpha/k)$

"Extra" increase in doping

"Smaller" reduction in supply voltage

Scaling of device parameters (k>1, α>1)

Electric field → α

Carrier velocity → α (or 1 for velocity saturation)

Depletion layer width → 1/k

Capacitance → 1/k

Inversion charge density → α

Drift current (on-state, per W) → α^2 (or α for velocity saturation)

Diffusion current (off-state, per W) → α.k

Scaling of circuit parameters

	conventional	(velocity saturation)
Delay time (C.V/I) →	$1/\alpha k$	(or $1/k$)
Power dissipation (V.I) →	α^3/k^2	(or α^2/k^2)
Power-delay product →	α^2/k^3	
Circuit density (1/Area) →	k^2	
Power density (P/A) →	α^3	(or α^2)