UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

The MOS(FET) Transistor

A.Carini – Digital Integrated Circuits

The MOS(FET) Transistor

- It is the workhorse of contemporary digital design.
- Its major assets from a digital perspective is that:
  - It performs very well as a switch,
  - It introduces little parasitic effects.
  - It allows a high integration density, and
  - Has a relatively “simple” manufacturing process.
- The MOSFET is a four terminal device.
- The voltage applied to the *gate* terminal determines if and how much current flows between the *source* and the *drain* ports.
- The *body* represents the fourth terminal, but its function is secondary as it only serves to modulate the device characteristics and parameters.

The MOS(FET) Transistor

- The transistor can be considered to be a switch.
- When a voltage, larger than a *threshold voltage* V_T, is applied to the gate, a conducting channel is formed between drain and source.
- In the presence of a voltage V_{DS}, current flows between drain and source.
- The conductivity of the channel is modulated by the gate voltage—the larger the voltage difference V_{GS}, the smaller the resistance of the conducting channel and the larger the current.
- When the gate voltage is lower than V_T, no such channel exists, and the switch is considered open.
- Two type of MOSFET: NMOS and PMOS.
- The NMOS consists of *n* doped drain and source regions, embedded in a *p*-type substrate. The current is carried by electrons moving through an *n*-type channel between S and D.
- For PMOS: *p* doped drain and source regions. Current of holes in a *p* channel.

The MOS(FET) Transistor

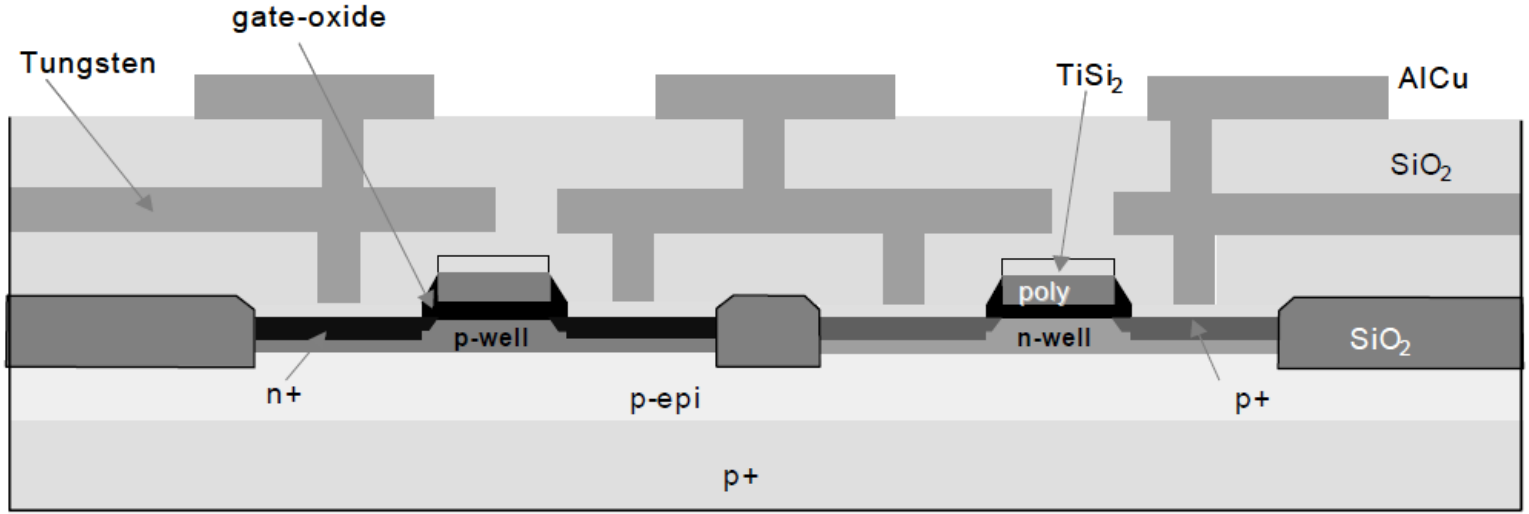
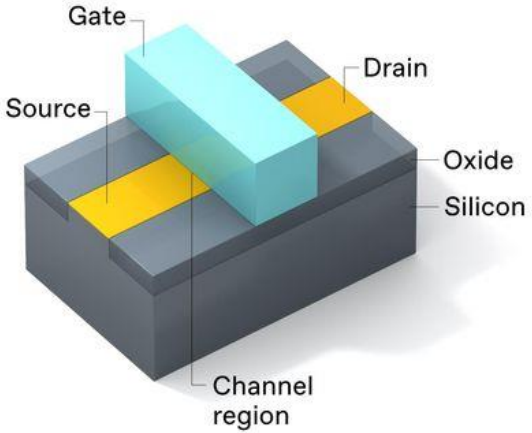
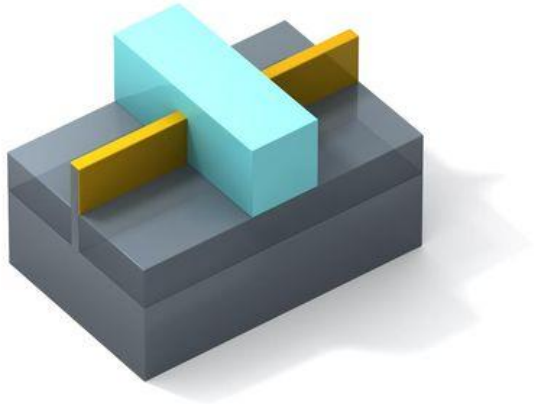


Figure 3.11 Cross-section of contemporary dual-well CMOS process. (fino al 2011)

The MOS(FET) Evolution

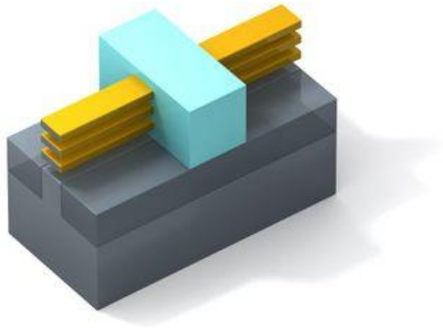




“The shift from a planar transistor architecture [left] to the FinFET [right] provided greater control of the channel [covered by blue box], resulting in a reduction in power consumption of 50 percent and an increase in performance of 37 percent. EMILY COOPER”
FROM: HTTPS://SPECTRUM.IEEE.ORG/3D-CMOS

It suppresses the "leaky" part of the silicon body, that is the regions where the electrostatic control of the gate is weak.
Can be implemented with the standard CMOS manufacturing processes.
Fabrication costs are comparable with planar technology.
Reduced body effect.

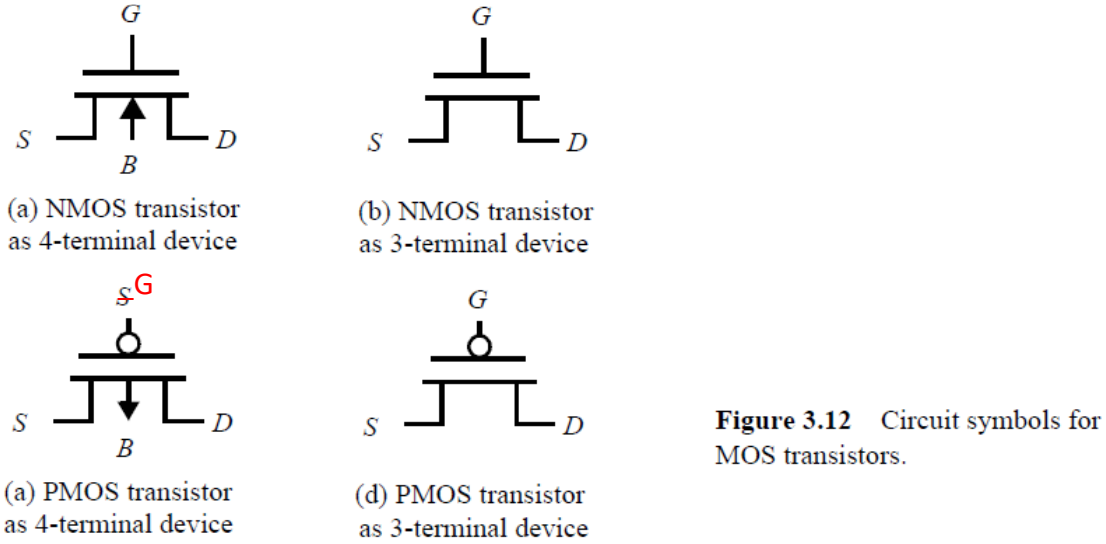
The MOS(FET) Evolution



In the RibbonFET, the gate wraps around the transistor channel region to enhance control of charge carriers. The new structure also enables better performance and more refined optimization. EMILY COOPER [because of a short channel effect mitigation]

FROM: HTTPS://SPECTRUM.IEEE.ORG/3D-CMOS

The MOS(FET) Transistor



(a) NMOS transistor as 4-terminal device

(b) NMOS transistor as 3-terminal device

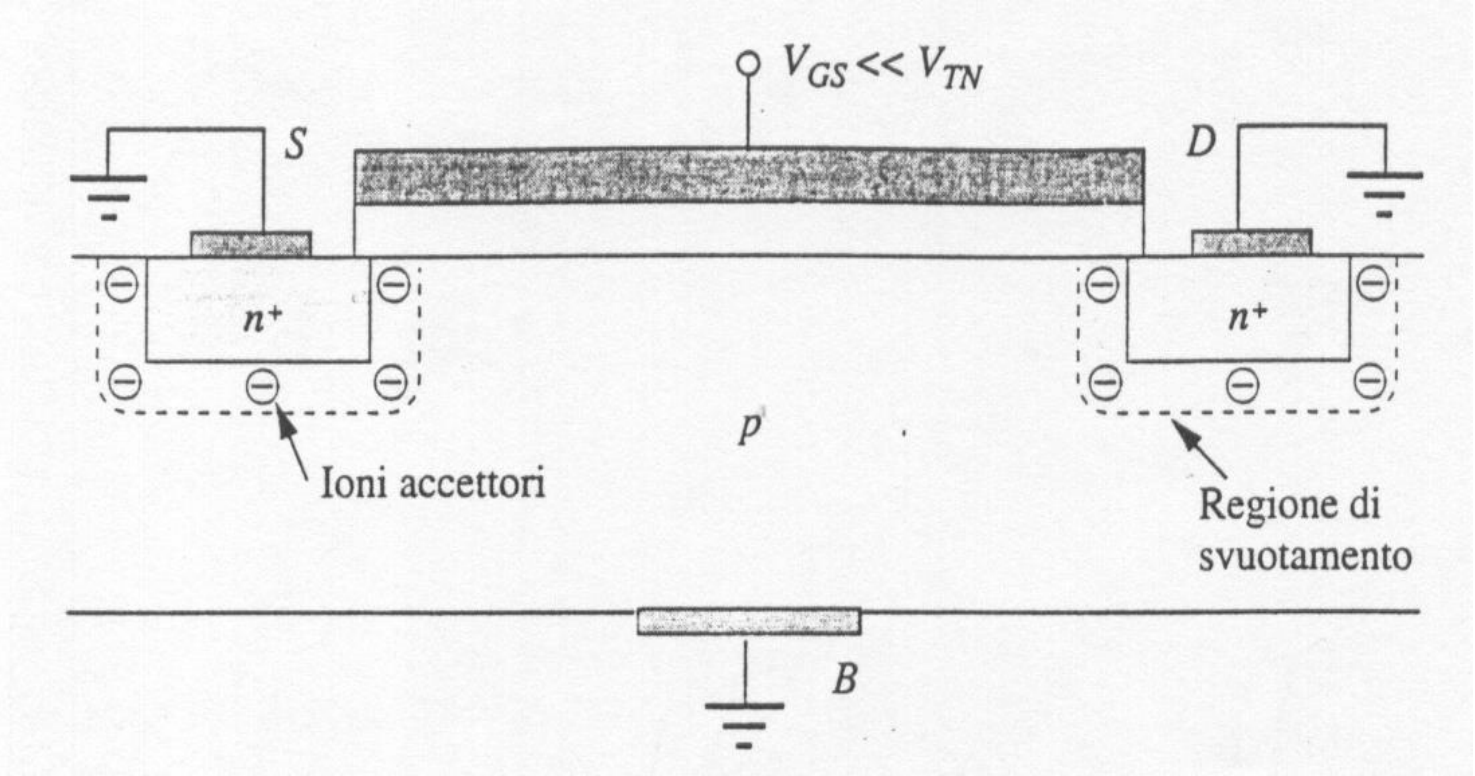
(a) PMOS transistor as 4-terminal device

(d) PMOS transistor as 3-terminal device

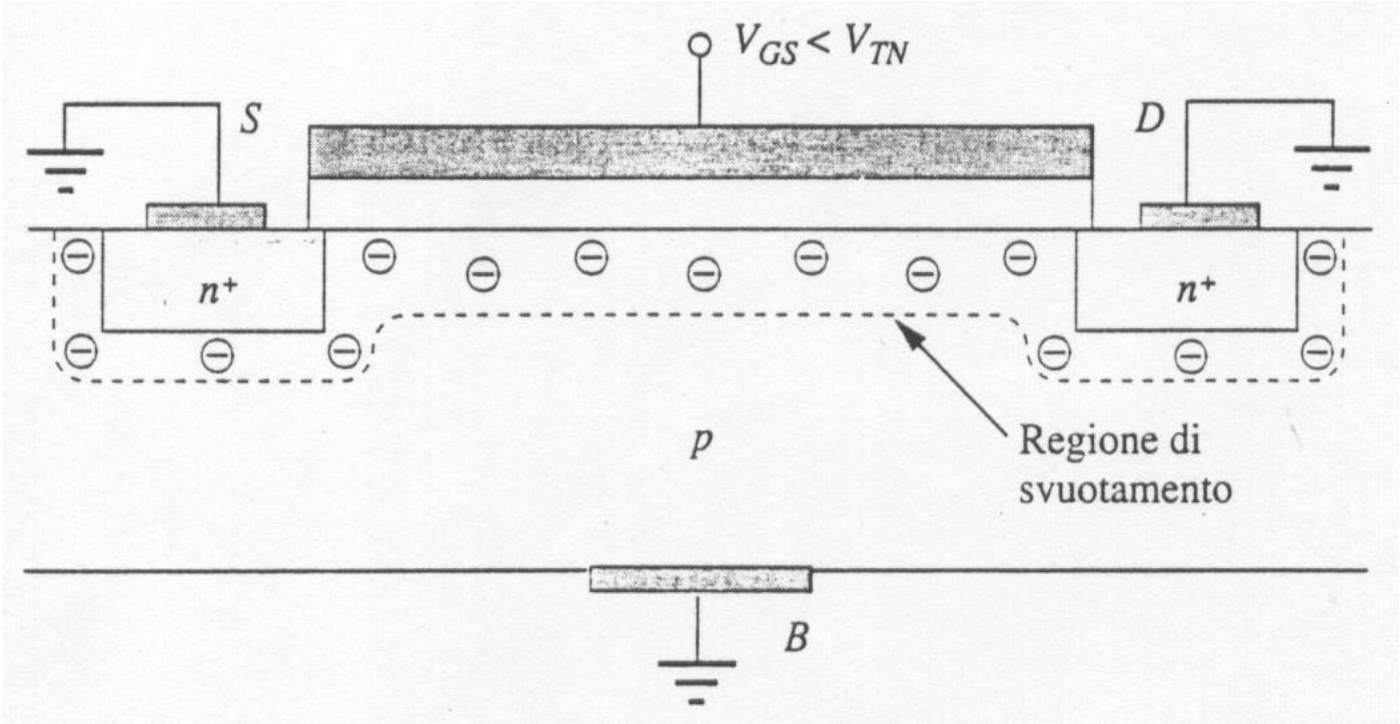
Figure 3.12 Circuit symbols for MOS transistors.

- If the fourth terminal is not shown, it is assumed that the body is connected to the appropriate supply.
- Source and drain are interchangeable. In NMOS the source is the lowest voltage terminal, in PMOS it is the highest voltage terminal.

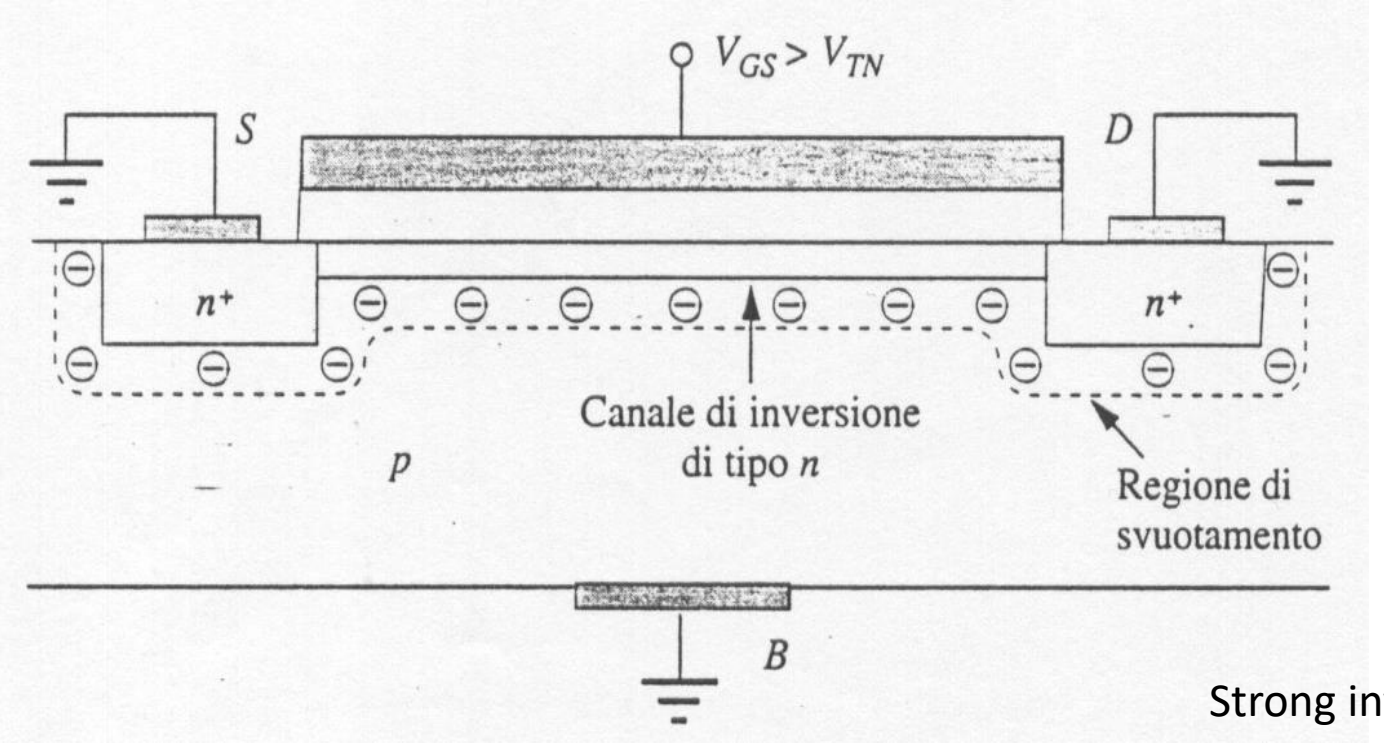
The MOS Transistor under Static Conditions



The MOS Transistor under Static Conditions



The MOS Transistor under Static Conditions



Strong inversion

The MOS Transistor under Static Conditions

- The value of V_{GS} where strong inversion occurs is called the *threshold voltage* V_T.
- V_T is a function of several components, e.g., the oxide thickness, the substrate dope level, the charge of impurities trapped at the surface between channel and gate oxide, and the dosage of ions implanted for threshold adjustment.
- The source-bulk voltage V_{SB} has an impact on the threshold:

$$V_T = V_{T0} + \gamma\left(\sqrt{\left|-2\phi_F + V_{SB}\right|} - \sqrt{\left|-2\phi_F\right|}\right)$$

- γ is called the *body-effect coefficient,* Φ_F is the Fermi potential.
- V_T and γ are positive for an NMOS, negative for a PMOS.
- $|V_{SB}|>0$ increases $|V_T|$

$$\phi_F = -\phi_T \ln\left(\frac{N_A}{n_i}\right) \qquad (\approx 0.3\ \text{V})$$

Resistive Operation

- Assume $V_{GS} > V_T$ and a small voltage, V_{DS}, is applied between drain and source.
- The voltage difference causes a current I_D to flow from drain to source.

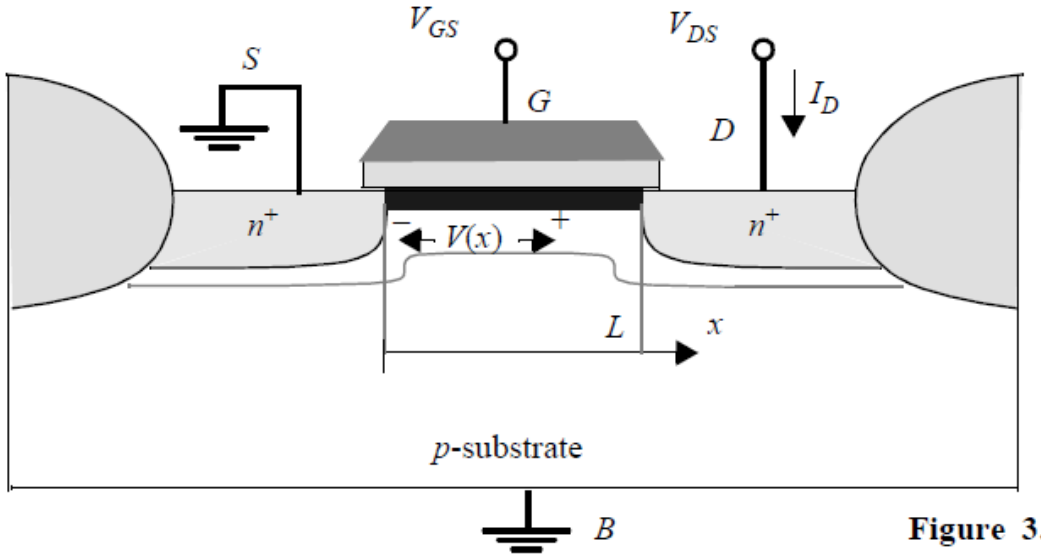


Figure 3.15 NMOS transistor with bias voltages.

Resistive Operation

- In the hypothesis that the drift velocity of charges is proportional to the electric field,

$$\upsilon_n = -\mu_n \xi(x) = \mu_n \frac{dV}{dx}$$

- where μ_n is the electron *mobility,* it can be proved that

$$I_D = k'_n \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] = k_n\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right]$$

- k'_n is called the *process transconductance parameter* and equals

$$k'_n = \mu_n C_{ox} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}}$$

- $\varepsilon_{ox} = 3.97 \times \varepsilon_o = 3.5 \times 10^{-11}$ F/m the oxide permittivity, and t_{ox} the oxide thickness.
- The product of k'_n and the *W/L* ratio is called the *gain factor* k_n of the device.

Resistive Operation

- For smaller values of V_{DS}, the quadratic factor can be ignored, and we observe a linear dependence between V_{DS} and I_D.
- The operation region is hence called the *resistive* or *linear* region.
- The *W* and *L* parameters represent the *effective channel width and length* of the transistor. These values differ from the dimensions *drawn* on the layout due to effects such as lateral diffusion of the source and drain regions.

The Saturation Region

- As the value of the drain-source voltage increases, the assumption that the channel voltage is larger than the threshold all along the channel ceases to hold.
- When $V_{DS} \geq V_{GS}-V_T$ the conducting channel disappears or is *pinched off.*

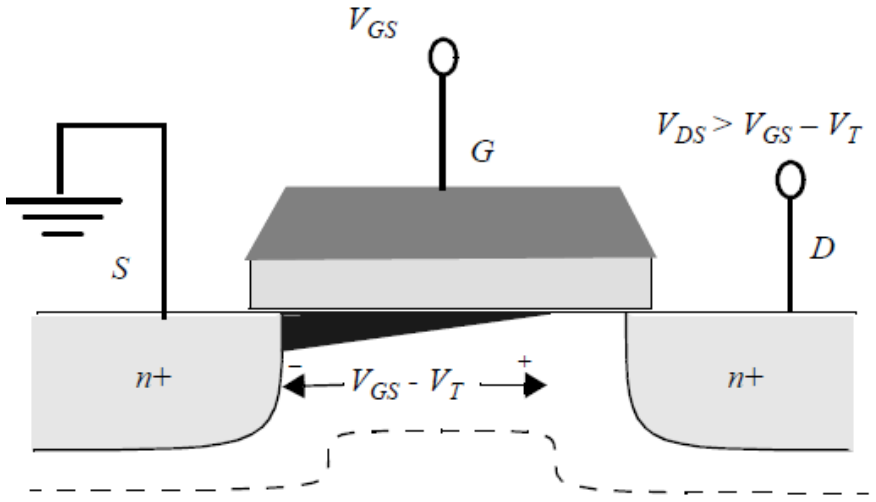


Figure 3.16 NMOS transistor under pinch-off conditions.

The Saturation Region

- The voltage difference over the induced channel (from pinch-off to source) remains fixed at $V_{GS} - V_T$, and thus the current remains constant (or saturates).
- Replacing V_{DS} by $V_{GS} - V_T$:

$$I_D = \frac{k'_n}{2}\frac{W}{L}(V_{GS} - V_T)^2$$

- To a first agree, the current is no longer a function of V_{DS}.
- In reality, the effective length of the conductive channel is modulated by V_{DS}.
- More accurately:

$$I_D = I_D'(1 + \lambda V_{DS})$$

- With λ an empirical parameter, called the *channel-length modulation*.
- λ varies roughly with the inverse of the channel length.
- In shorter transistors, the pinch-off region presents a larger fraction of the channel, and the channel-modulation effect is more pronounced.

Velocity Saturation

- The behavior of transistors with very short channel lengths (called *short-channel devices*) deviates considerably from the resistive and saturated models.
- The main culprit for this deficiency is the *velocity saturation* effect.
- At high field strengths, the carrier velocity isn't proportional to the electric field.
- When the electrical field along the channel reaches a critical value ξ_c, the velocity tends to saturate due to scattering effects (collisions suffered by the carriers).

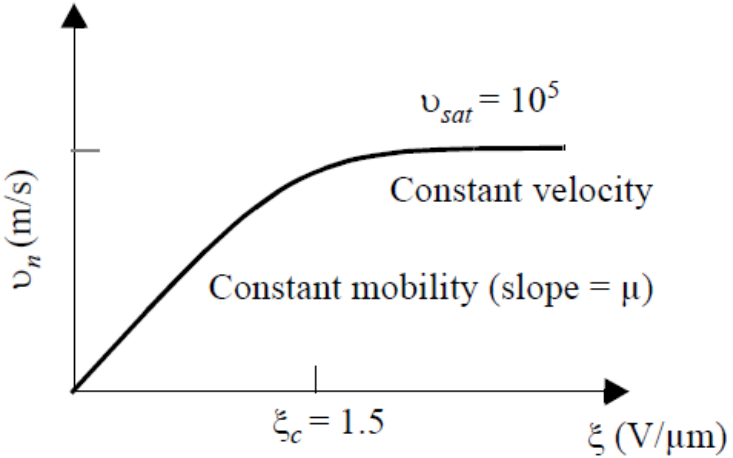


Figure 3.17 Velocity-saturation effect.

Velocity Saturation

- The saturation velocity of electrons and holes approximately equals 10^5 m/s.
- The critical field is around 1.5 V/um for electrons, but is higher for holes.
- Velocity-saturation effects are hence less pronounced in PMOS transistors.
- For short-channel devices, because of the velocity saturation the delivered current is smaller than what would be normally expected.
- For a short-channel device and for large enough values of $V_{GT}=V_{GS}-V_T$, $\boldsymbol{V_{DSAT} < V_{GT}}$. The device enters saturation before V_{DS} reaches $V_{GS} - V_T$.
- The saturation current $\boldsymbol{I_{DSAT}}$ **displays a *linear dependence*** with respect to $\boldsymbol{V_{GS}}$ voltage, in contrast with the squared dependence in the long channel devices.
- This reduces the amount of current a transistor can deliver for a given V_{GS}.

Velocity Saturation – a simple model

- A simple model can be obtained by making two assumptions:
  1. The velocity saturates abruptly at ξ_c, and is approximated by

$$\upsilon = \mu_n \xi \quad \text{for} \quad \xi \le \xi_c$$
$$= \upsilon_{sat} = \mu_n \xi_c \quad \text{for} \quad \xi \ge \xi_c$$

  2. The drain-source voltage V_{DSAT} at which the critical electrical field is reached and velocity saturation comes into play is *constant* and is approximated by

$$V_{DSAT} = L\xi_c = \frac{L\upsilon_{sat}}{\mu_n}$$

- Under these circumstances, the current equations for the resistive region remain unchanged from the long-channel model.
- Once V_{DSAT} is reached, the current abruptly saturates.

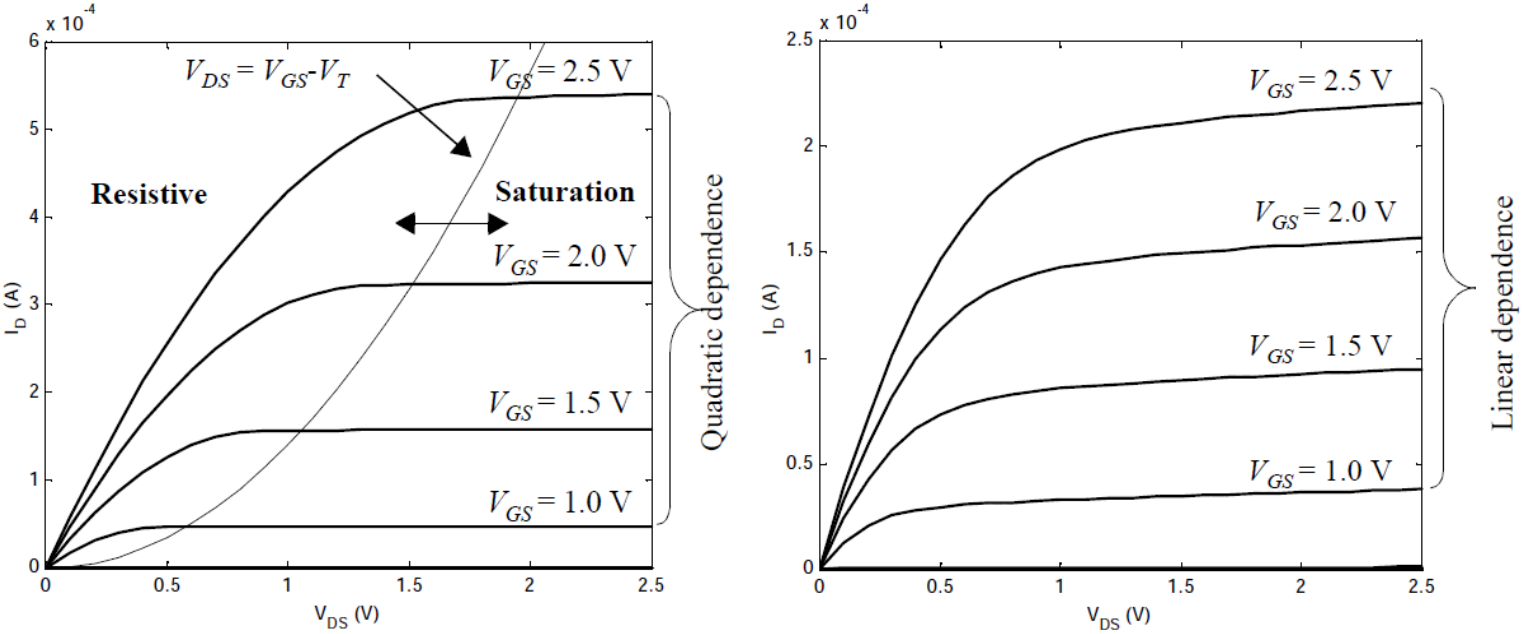
Velocity Saturation – a simple model

$$I_{DSAT} = I_D(V_{DS} = V_{DSAT})$$

$$= \mu_n C_{ox} \frac{W}{L}\left((V_{GS} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}\right)$$

$$= \upsilon_{sat} C_{ox} W\left(V_{GS} - V_T - \frac{V_{DSAT}}{2}\right)$$

Drain Current versus Voltage Charts

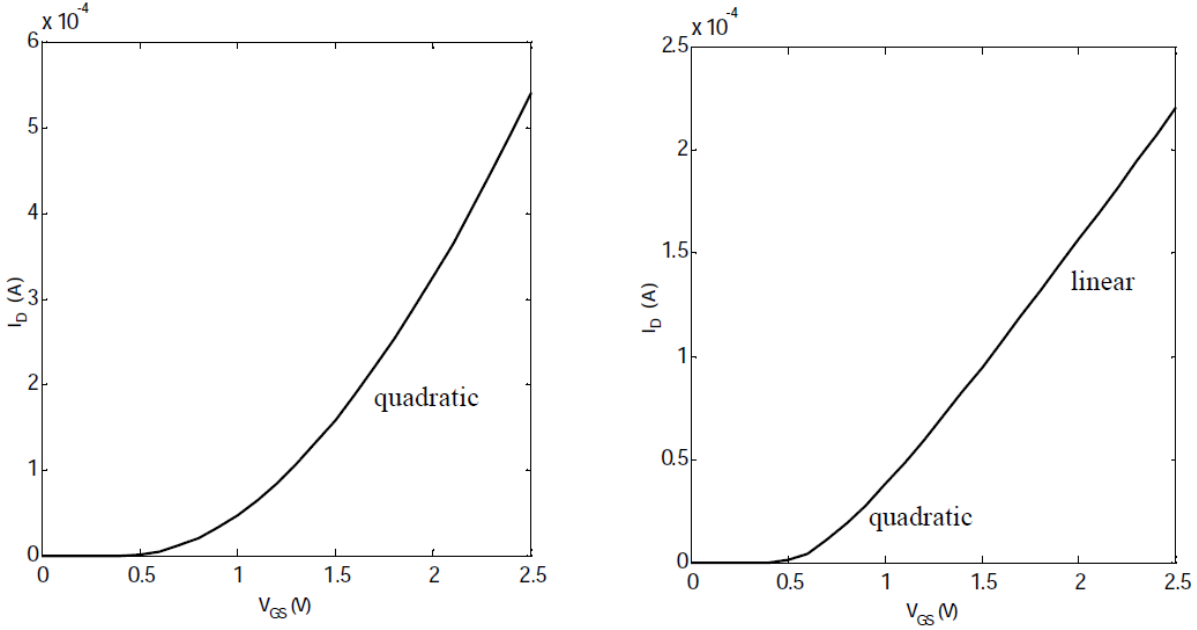


(a) Long-channel transistor ($L_d = 10$ μm)

(b) Short-channel transistor ($L_d = 0.25$ μm)

Figure 3.19 *I-V* characteristics of long- and a short-channel NMOS transistors in a 0.25 μm CMOS technology. The (*W*/*L*) ration of both transistors is identical and equals 1.5

Drain Current versus Voltage Charts



(a) Long-channel device (L_d = 10 μm)

(b) Short-channel device (L_d = 0.25 μm)

Figure 3.20 NMOS transistor I_D-V_{GS} characteristic for long and short-channel devices (0.25 μm CMOS technology). W/L = 1.5 for both transistors and V_{DS} = 2.5 V.

Subthreshold Conduction

- The current does not drop abruptly to 0 at $V_{GS} = V_T$.
- It becomes apparent that the MOS transistor is already partially conducting for voltages below the threshold voltage.
- This effect is called *subthreshold* or *weak-inversion* conduction.
- The current for $V_{GS} < V_T$ can be approximated by the expression

$$I_D = I_S e^{\frac{V_{GS}}{nkT/q}} \left(1 - e^{-\frac{V_{DS}}{kT/q}}\right)$$

- where I_S and n are empirical parameters, with $n \geq 1$ and typically around 1.5.

Subthreshold Conduction

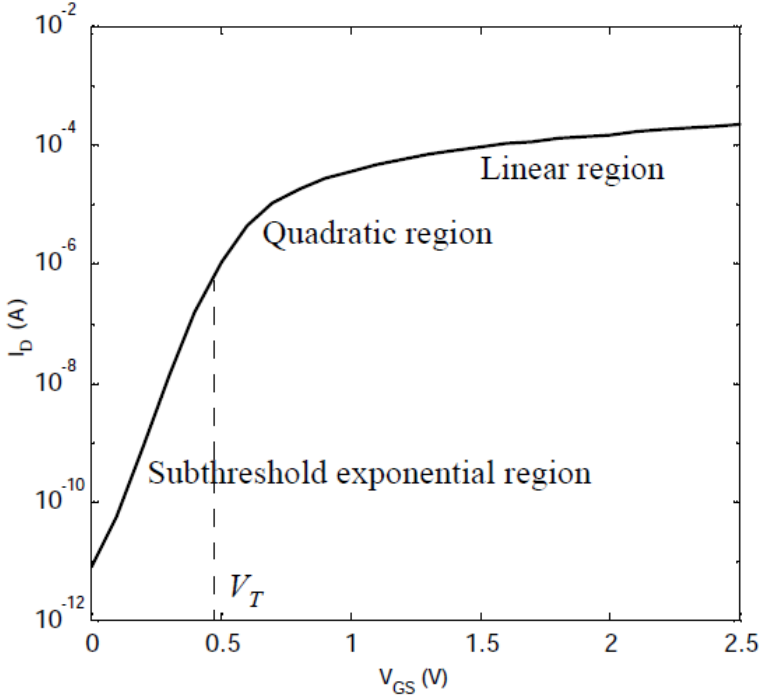


Figure 3.22 I_D current versus V_{GS} (on logarithmic scale), showing the exponential characteristic of the subthreshold region.

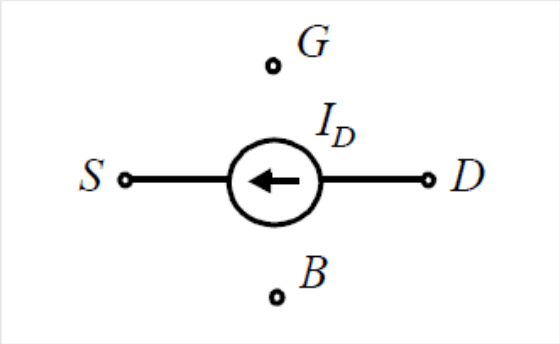
Subthreshold Conduction

- In most digital applications, the presence of subthreshold current is undesirable.
- We would rather have the current drop as fast as possible once the gate-source voltage falls below V_T.
- The (inverse) rate of decline of the current with respect to V_{GS} below V_T hence is a quality measure of a device.
- It is often quantified by the *slope factor S,* which measures by how much V_{GS} has to be reduced for the drain current to drop by a factor of 10.

$$S = n\left(\frac{kT}{q}\right)\ln(10)$$

- For an ideal transistor with the sharpest possible roll-off, $n = 1$ and $(kT/q)\ln(10)$ evaluates to 60 mV/decade at room temperature.

A unified MOS model for manual analysis



$$I_D = 0 \quad \text{for} \quad V_{GT} \le 0$$

$$I_D = k' \frac{W}{L}\left(V_{GT} V_{min} - \frac{V_{min}^2}{2} \right)(1 + \lambda V_{DS}) \quad \text{for} \quad V_{GT} \ge 0$$

with $V_{min} = \min(V_{GT}, V_{DS}, V_{DSAT})$,

$V_{GT} = V_{GS} - V_T$,

and $V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|})$

A unified MOS model for manual analysis

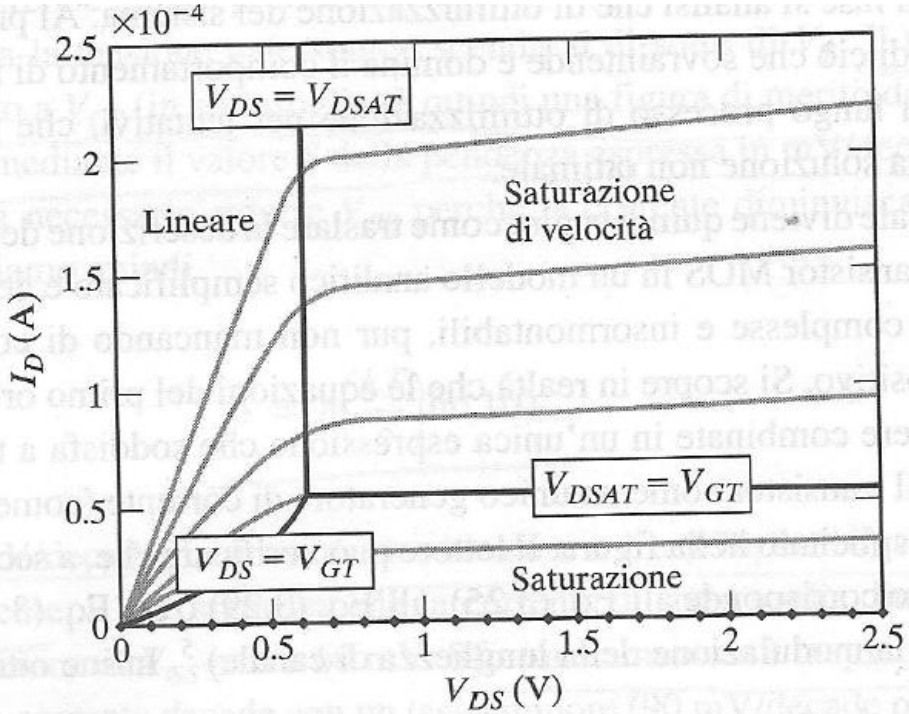


Figura 3-24 Limiti delle regioni operative, in accordo al modello unificato, per l'analisi manuale.

A unified MOS model for manual analysis

Table 3.2 Parameters for manual model of generic 0.25 μm CMOS process (minimum length device).

	V_{T0} (V)	γ ($V^{0.5}$)	V_{DSAT} (V)	k' (A/V^2)	λ (V^{-1})
NMOS	0.43	0.4	0.63	115×10^{-6}	0.06
PMOS	–0.4	-0.4	-1	-30×10^{-6}	-0.1

An even more simplified model

- We introduce an even more simplified model that has the advantage of being linear and straightforward.
- It is based on the underlying assumption that the transistor is nothing more than a switch with an infinite off-resistance, and a finite on-resistance R_{on}.

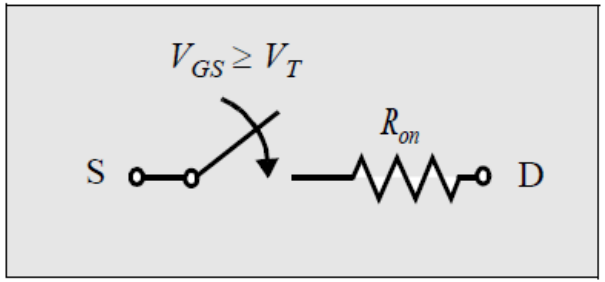


Figure 3.25 NMOS transistor modeled as a switch.

An even more simplified model

- R_{on} is still time-variant, non-linear and depending upon the operation point of the transistor.
- When studying digital circuits in the transient mode it is attractive to assume R_{on} as a **constant** and **linear** resistance R_{eq}, chosen so that the final result is similar to what would be obtained with the original transistor.
- A reasonable approach in that respect is to use the **average value** of the resistance **over the operation region** of interest, or even simpler, the **average value** of the resistances **at the end-points** of the transition.

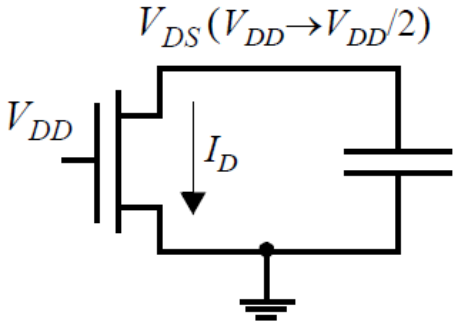
$$R_{eq} = \text{average}_{t=t_1 \ldots t_2}(R_{on}(t)) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} R_{on}(t)dt = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{V_{DS}(t)}{I_D(t)}dt$$

$$\approx \frac{1}{2}(R_{on}(t_1) + R_{on}(t_2))$$

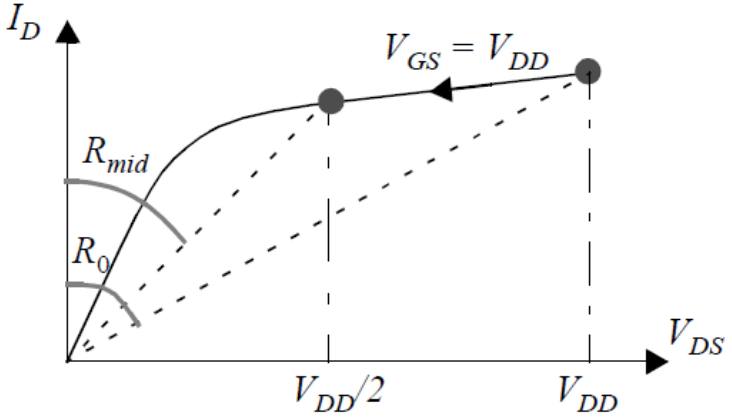
Equivalent resistance when (dis)charging a capacitor

- One of the most common scenario's in contemporary digital circuits is the discharging of a capacitor from V_{DD} to GND through an NMOS transistor with its gate voltage set to V_{DD}, or vice-versa the charging of the capacitor to V_{DD} through a PMOS with its gate at GND.
- Of special interest is the point where the voltage on the capacitor reaches the mid-point ($V_{DD}/2$) by virtue of the definition of the propagation delay.
- Assuming that the supply voltage is substantially larger than the velocity-saturation voltage V_{DSAT} of the transistor, it is fair to state that the transistor stays in velocity saturation for the entire duration of the transition.

Equivalent resistance when (dis)charging a capacitor



(a) schematic



(b) trajectory traversed on ID-VDS curve.

$$R_{eq} = \frac{1}{-V_{DD}/2}\int_{V_{DD}}^{V_{DD}/2} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right) \quad \longleftarrow \quad \frac{dT}{T} \approx \frac{dV}{-V_{DD}/2}$$

$$\text{with} \quad I_{DSAT} = k'\frac{W}{L}\left((V_{DD}-V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}\right)$$

Equivalent resistance when (dis)charging a capacitor

- A similar result can be obtained by just averaging the values of the resistance at the end points (and simplifying the result using a Taylor expansion):

$$R_{eq} = \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT}(1+\lambda V_{DD})} + \frac{V_{DD}/2}{I_{DSAT}(1+\lambda V_{DD}/2)}\right) \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{5}{6}\lambda V_{DD}\right)$$

- We can notice that:
  - The resistance is inversely proportional to the (W/L) ratio of the device.
  - For $V_{DD} >> V_T + V_{DSAT}/2$, the resistance becomes virtually independent of V_{DD}.
  - Once the supply voltage approaches V_T, a dramatic increase in resistance can be observed.

Equivalent resistance when (dis)charging a capacitor

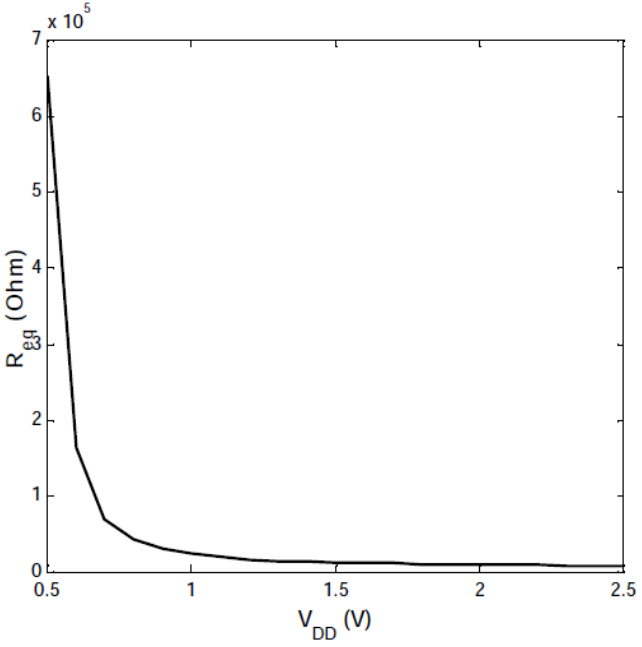
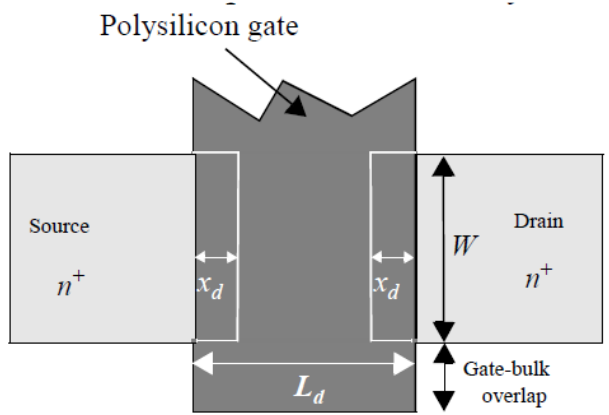


Figure 3.27 Simulated equivalent resistance of a minimum size NMOS transistor in 0.25 μm CMOS process as a function of V_{DD} ($V_{GS} = V_{DD}$, $V_{DS} = V_{DD} \rightarrow V_{DD}/2$).

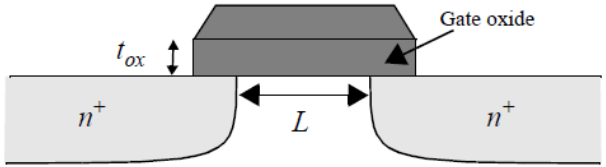
Dynamic Behavior

- The dynamic response of a MOSFET transistor is a sole function of the time it takes to (dis)charge the parasitic capacitances that are intrinsic to the device, and the extra capacitance introduced by the interconnecting lines.
- The intrinsic capacitances originate from three sources:
  - the basic MOS structure,
  - the channel charge, and
  - the depletion regions of the reverse-biased *pn*-junctions of drain and source.
- Aside from the MOS structure capacitances, all capacitors are nonlinear and vary with the applied voltage.

MOS Structure Capacitances



(a) Top view



(b) Cross section

$$C_{GSO} = C_{GDO} = C_{ox}x_dW = C_oW$$

Figure 3.28 MOSFET overlap capacitance.

Channel Capacitance

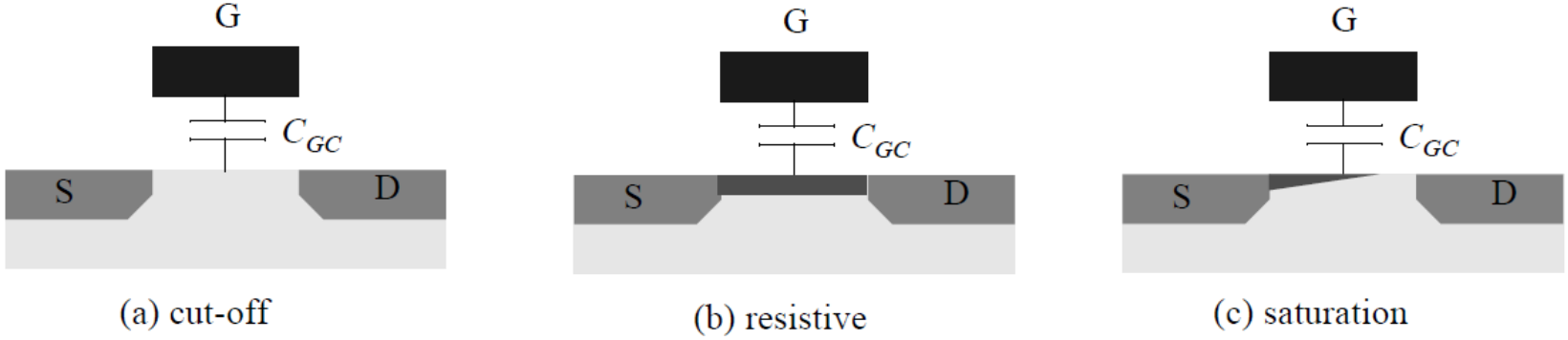


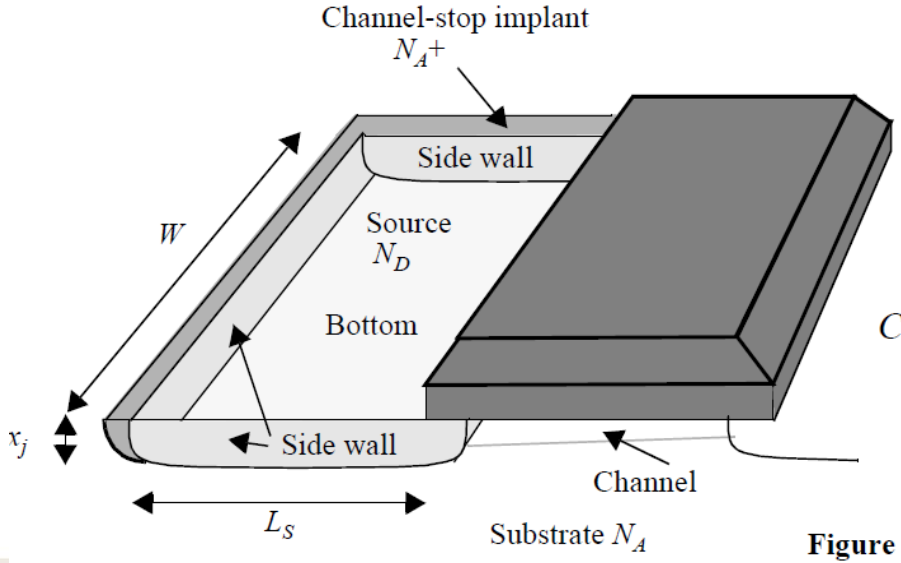
Figure 3.29 The gate-to-channel capacitance and how the operation region influences is distribution over the three other device terminals.

Table 3.4 Average distribution of channel capacitance of MOS transistor for different operation regions.

Operation Region	C_{GCB}	C_{GCS}	C_{GCD}	C_{GC}	C_G
Cutoff	$C_{ox}WL$	0	0	$C_{ox}WL$	$C_{ox}WL+2C_oW$
Resistive	0	$C_{ox}WL/2$	$C_{ox}WL/2$	$C_{ox}WL$	$C_{ox}WL+2C_oW$
Saturation	0	$(2/3)C_{ox}WL$	0	$(2/3)C_{ox}WL$	$(2/3)C_{ox}WL+2C_oW$

Junction Capacitances

- A capacitive component is contributed by the reverse-biased source-body and drain-body *pn*-junctions.
- The depletion-region capacitance is nonlinear and decreases when the reverse bias is raised.



$$C_{diff} = C_{bottom} + C_{sw} = C_j \times AREA + C_{jsw} \times PERIMETER$$
$$= C_j L_S W + C_{jsw}(2L_S + W)$$

Figure 3.32 Detailed view of source junction.

Capacitive Device Model

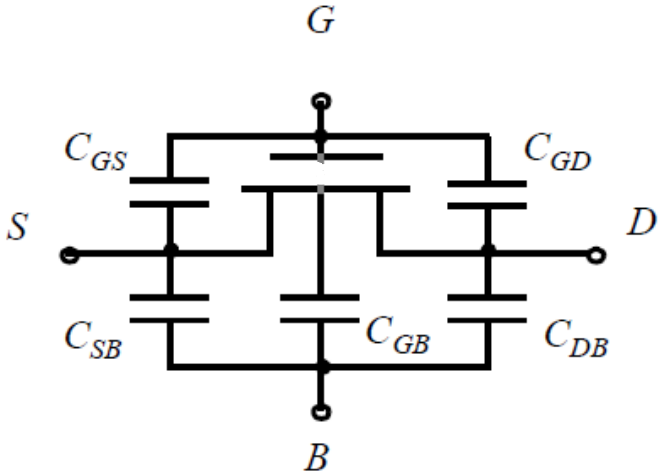


Figure 3.33 MOSFET capacitance model.

$$C_{GS} = C_{GCS} + C_{GSO};\ C_{GD} = C_{GCD} + C_{GDO};\ C_{GB} = C_{GCB}$$

$$C_{SB} = C_{Sdiff};\ C_{DB} = C_{Ddiff}$$

See:

- J. M. Rabaey, A. Chandrakasan, B. Nikolic, «Digital Integrated Circuits: A Design Perspective», Pearson, 2003
  - Cap. 3.3