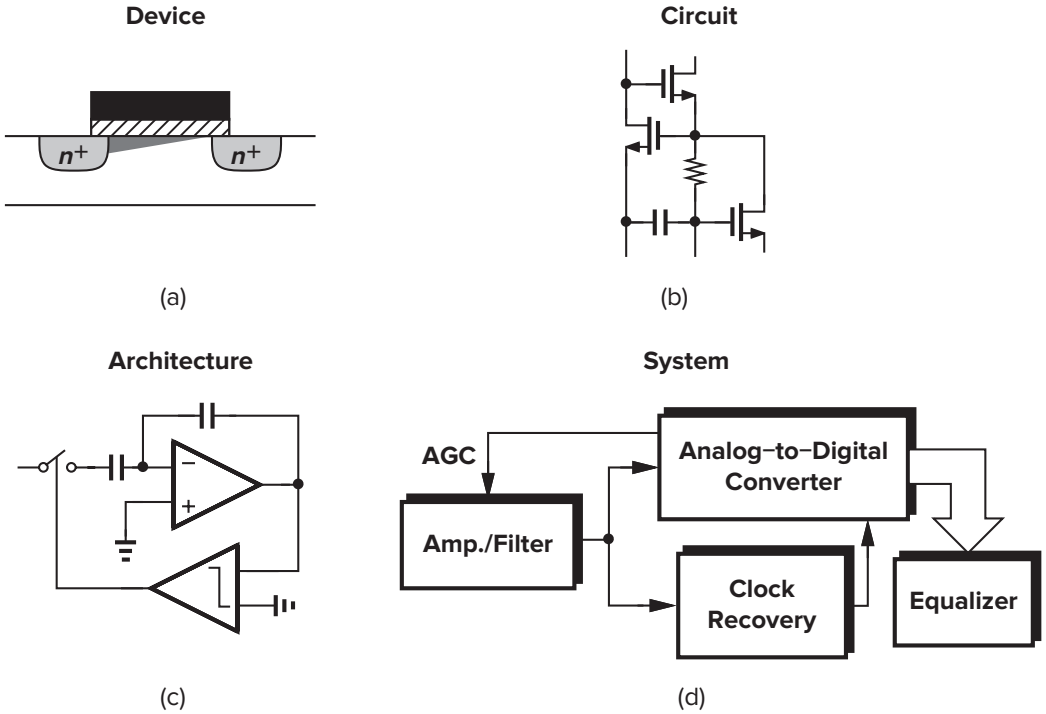
Notes for *Design of Analog CMOS Integrated Circuits*

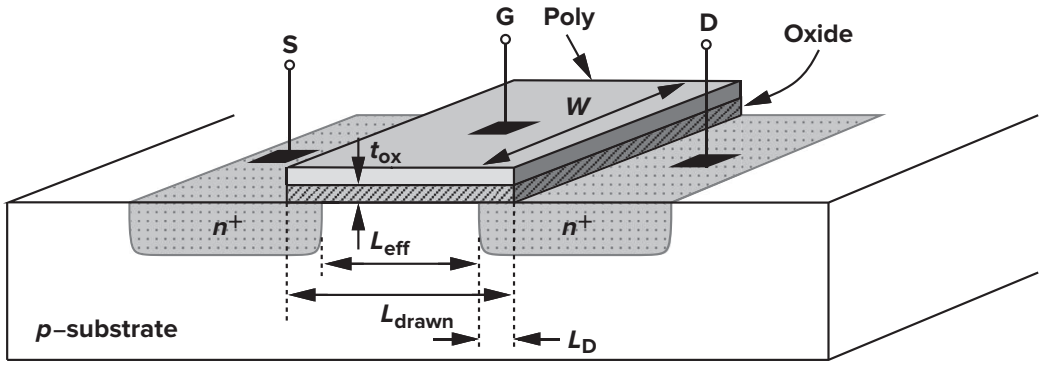
CMOS 模拟集成电路设计笔记

Yi Ding
(University of Chinese Academy of Sciences, Beijing 100049, China)

丁毅
(中国科学院大学，北京 100049)

2024.11 – ...





Preface

to be completed

Table 1: Learning Plan

Task	Week	Date	Planned Pages (actual)
☑	1	2025.01.13 - 2025.01.19	7 - 46 ()
	2	2025.01.20 - 2025.01.26	47 -
	3	2025.01.27 - 2025.02.02	87 -
	4	2025.02.03 - 2025.02.09	127 -
	5	2025.02.10 - 2025.02.16	167 -
	6	2025.02.17 - 2025.02.23	207 -

序言

待完成

Contents

Chapter 1 Introduction to Analog Design

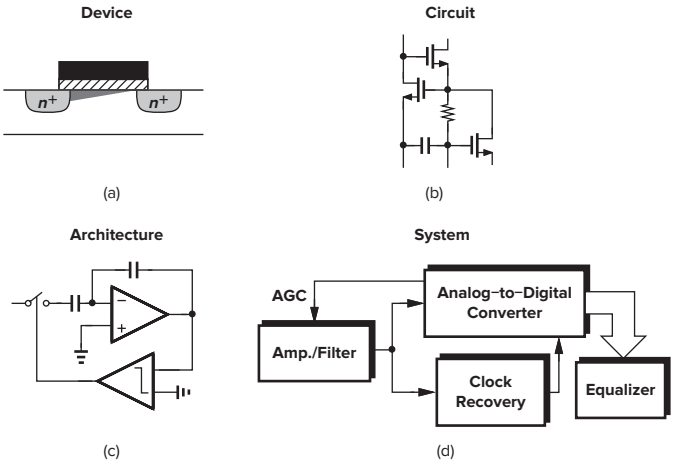


Figure 1.1: Abstraction levels in analog design:
(a) device level;
(b) circuit level;
(c) architecture level;
(d) system level

Chapter 2 Basic MOS Device Physics

In this chapter, we study the physics of MOSFETs at an elementary level, covering the bare minimum that is necessary for basic analog design. The ultimate goal is still to develop a circuit model for each device by formulating its operation.

After studying many analog circuits in Chapters 3 through 14 and gaining motivation for a deeper understanding of devices, we return to the subject in Chapter 17 and deal with other aspects of MOS operation including more advanced properties and second-order effects.

2.1 General Considerations

Figure 2.1 shows a simplified structure of an n-type MOSFET (NMOS) device.

(1) p-type substrate: also called **bulk** or **body**;
(2) a heavily-doped (conductive) piece of polysilicon (called **poly**) operating as the gate;
(3) a thin layer of silicon dioxide (SiO_2) (called **oxide**) insulats the gate from the substrate;
(4) $L_{\text{eff}} = L_{\text{drawn}} - 2L_D$, where L_{eff}[①] is the effective channel length (typically 10 nm in 2015), L_{drawn} is the total length, and L_D is the amount of side diffusion.
(5) t_{ox}: gate oxide thickness (typical 15 Å in 2015)

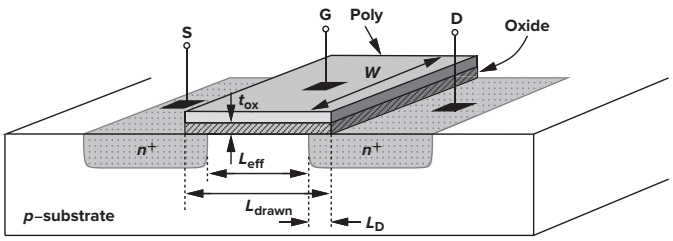


Figure 2.1: Simplified structure of an NMOS device

In practice, NMOS and PMOS devices must be fabricated on the same wafer, i.e., the same substrate. For this reason, one device type can be placed in a **local substrate**, usually called a **well**. In today's CMOS processes, the PMOS device is fabricated in an n-well (on the p-type substrate), depicted in Figure 2.2.

Figure 2.2 indicates that, while all NFETs share the same substrate, each PFET can have an independent n-well. This flexibility of PFETs is exploited in some analog circuits.

> Some modern CMOS processes offer a **deep n-well** (an n-well that contains an NMOS device and its p-type bulk), so that the NMOS device can be isolated from the other NMOS devices.

The circuit symbols used to represent NMOS and PMOS transistors are shown in Figure 2.3. The letter "B" stands for **bulk** or **body**, i.e., the substrate of the device. The source of the PMOS device is positioned on top as a visual aid because it has a higher potential than its gate.

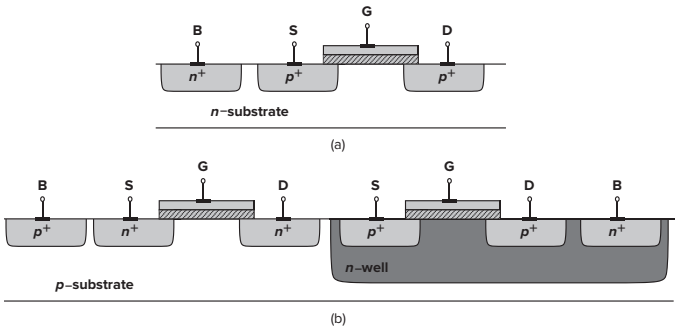


Figure 2.2: CMOS processes. (a) A simple PMOS device; (b) NMOS and PMOS devices on the same substrate

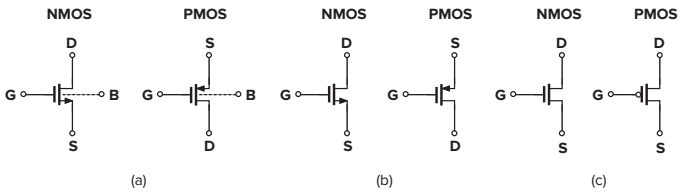


Figure 2.3: Circuit symbols for MOSFETs.
(a) NMOS and PMOS devices;
(b) omit bulk connections (GND for NMOS and VDD for PMOS);
(c) digital representation of NMOS and PMOS devices

In this book, we prefer those in Figure 2.3(b) to gain a clear view of the device.

2.2 MOS I/V Characteristics

2.2.1 Threshold Voltage V_{TH}

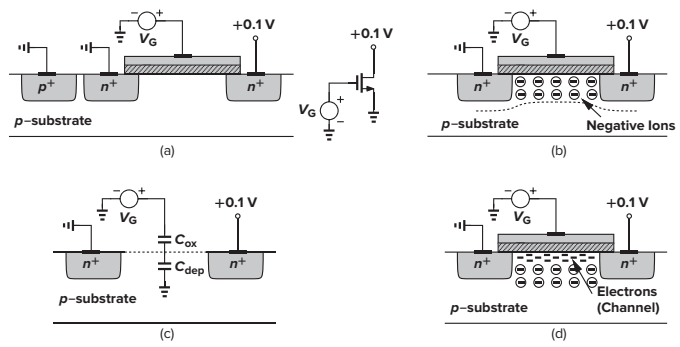


Figure 2.4: Formation of inversion layer in an NMOS.
(a) A MOSFET driven by a gate voltage;
(b) formation of depletion region;
(c) onset of inversion;
(d) formation of inversion layer.

[①] In the remainder of this book, we denote the effective length L_{eff} by L unless otherwise stated.

When the interface potential reaches a sufficiently positive value (V_{GS}), a **channel** of charge carriers is formed under the gate oxide between S and D, and the transistor is "turned on". We say the interface is **inverted**, and the channel is called **inversion layer**.

In reality, the turn-on phenomenon is a gradual function of the gate voltage, making it difficult to define V_{TH} unambiguously. In semiconductor physics, the V_{TH} of an NFET is usually defined as the gate voltage for which the interface is "as much n-type as the substrate is p-type", given by

$$V_{TH} = \Phi_{MS} + 2\Phi_F + \frac{Q_{\text{dep}}}{C_{ox}} \tag{2.1}$$

where

(1) Φ_{MS}: work function difference between the polysilicon gate and the silicon substrate;
(2) Q_{dep}: the charge in the depletion region (see Figure 2.4);
(3) C_{ox}: the oxide capacitance per unit;
(4) Φ_F: Fermi potential difference between the surface and bulk, given by $\Phi_F = \frac{kT}{q_e} \ln \frac{N_{\text{sub}}}{n_i}$, where N_{sub} is the doping density of the substrate, n_i is the density of electrons in intrinsic silicon, and q_e is the electron charge.

Since C_{ox} appears frequently in device and circuit equations, it could be helpful to remember $C_{ox} \approx 17.25\ \text{fF}/\mu\text{m}^2$ for oxide thickness $t_{ox} \approx 20$ Å (see Figure 2.1). For other t_{ox}, bear in mind that $t_{ox}C_{ox} = 345\ \text{Å} \cdot \text{fF}/\mu\text{m}^2$ remains constant.

In practice, V_{TH} is usually adjusted by implantation of dopants into the channel area during device fabrication. In essence, to deplete the layer, if a thin layer of p^+ (n^-) is created, the threshold voltage of an NMOS device increases (decreases), depicted in Figure 2.5.

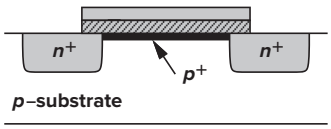


Figure 2.5: Implantation of p^+ dopants in an NMOS to alter (increase) the threshold voltage.

2.2.2 I/V Characteristics

Now considering an NMOS with $V_{GS} \geqslant V_{TH}$ (see Figure 2.6), and assume $Q_d = \frac{\mathrm{d}\rho_{\text{charge}}}{\mathrm{d}L} = \frac{\rho_{\text{charge}}}{L}$ is the mobile charge density along the direction of current, regard gate-channel as a capacitor, yielding

$$Q_d = WC_{ox}\left[V_{GS} - V_{TH} - V(x)\right] \tag{2.2}$$

where $V_(x)$ is the channel potential at x. The current I_D is given by

$$I_D = -Q_d \cdot v = -WC_{ox}\left[V_{GS} - V_{TH} - V(x)\right] \cdot \mu_n \frac{\mathrm{d}V(x)}{\mathrm{d}x}$$

subject to $V(0) = 0$ and $V(L) = V_{DS}$. Performing integration yields I_D as a function of V_{DS}

$$I_D = \mu_n C_{ox} \frac{W}{L}\left[\left(V_{GS} - V_{TH}\right)V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{2.3}$$

$$I_{D,\max} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}\left(V_{GS} - V_{TH}\right)^2 \tag{2.4}$$

Note that L is the effective channel length. We call $(V_{GS} - V_{TH})$ the **overdrive voltage** and $\frac{W}{L}$ the **aspect ratio**. We say the device is in the **triode region** (or linear region) if $V_{DS} \leqslant V_{GS} - V_{TH}$, and in the **saturation region** if $V_{DS} > V_{GS} - V_{TH}$.

Remark that the integration in (2.3) assumes that μ_n and V_{TH} are independent of x, V_D and V_G.

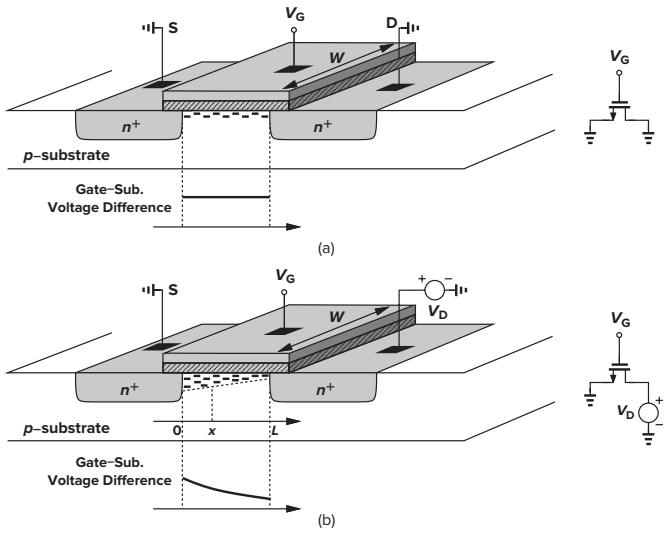


Figure 2.6: I/V characteristics of an NMOS device.
(a) $V_D = V_S$;
(b) $V_D > V_S$

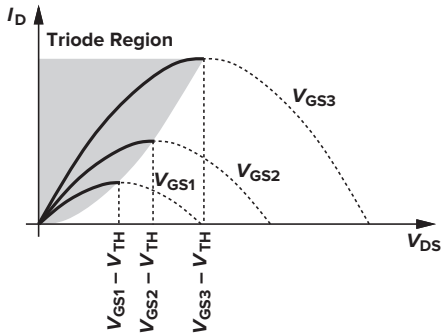


Figure 2.7: Triode region of an NMOS device.

In the deep triode region, i.e., $V_{DS} \ll 2(V_{GS} - V_{TH})$, we have

$$I_D \approx \mu_n C_{ox} \frac{W}{L}\left(V_{GS} - V_{TH}\right)V_{DS} \tag{2.5}$$

$$R_{\text{on}}|_{V_{DS}\to 0^+} = \frac{1}{\mu_n C_{ox}\frac{W}{L}\left(V_{GS} - V_{TH}\right)} \tag{2.6}$$

Therefore, as long as $V_{DS} \ll 2(V_{GS} - V_{TH})$, a MOSFET can operate as a voltage-controlled resistor [actually for $|V_{DS}| \ll 2(V_{GS} - V_{TH})$].

In the satyration region [$V_{DS} > (V_{GS} - V_{TH})$], the channel is **pinched off** (the inversion layer stops at $x < L$), leading to a relatively constant current I_D with respect to V_{DS}, depicted in Figure 2.6 and given by

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L'}(V_{GS} - V_{TH})^2 \tag{2.7}$$

$$V_{GS} = V_{TH} + \sqrt{\frac{2I_D}{\mu_n C_{ox}\frac{W}{L'}}} \tag{2.8}$$

We say the device exhibits a **square-law** behavior. As the electrons approach the pinch-off point where $Q_d \to 0$, their velocity rises tremendously ($v = \frac{I}{Q_d}$) so that they simply shoot through the depletion region and arrive at the drain terminal.

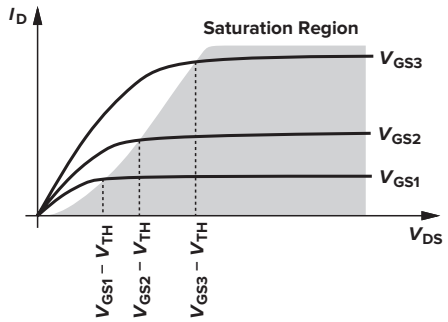


Figure 2.8: Saturation of drain current

For PMOS devices, the equations (2.3) and (2.7) are respectively written as

$$I_D = -\mu_p C_{ox}\frac{W}{L}\left[(V_{GS} - V_{TH})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{2.9}$$

$$I_D = -\frac{1}{2}\mu_p C_{ox}\frac{W}{L'}(V_{GS} - V_{TH})^2 \tag{2.10}$$

Note that V_{GS}, V_{DS}, V_{TH}, and $(V_{GS} - V_{TH})$ are negative for a PMOS transistor that is turned on. We can also rewrite the equations as

$$I_{SD} = \mu_p C_{ox}\frac{W}{L}\left[(V_{SG} - |V_{TH}|)V_{SD} - \frac{V_{SD}^2}{2}\right] \tag{2.11}$$

$$I_{SD} = \frac{1}{2}\mu_p C_{ox}\frac{W}{L'}(V_{SG} - |V_{TH}|)^2 \tag{2.12}$$

2.2.3 MOS Transconductance g_m

Define the transconductance g_m as the change in I_D with respect to V_{GS}, that is

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{2.13}$$

In the triode region, we have

$$g_m = \mu_n C_{ox}\frac{W}{L}V_{DS} \tag{2.14}$$

In the saturation region

$$g_m = \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH}) \tag{2.15}$$

$$= \sqrt{2\mu_n C_{ox}\frac{W}{L}I_D} \tag{2.16}$$

$$= \frac{2I_D}{V_{GS} - V_{TH}} \tag{2.17}$$

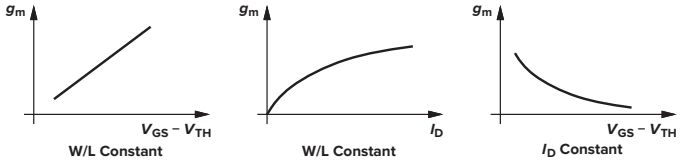


Figure 2.9: Approximate MOS transconductance as a function of overdrive and drain current.

Simply add a negative sign to obtain the transconductance of a PMOS device.

2.3 Second-Order Effects

2.3.1 Body Effect

Our analysis has so far entailed the assumption that the substrate (bulk, body) is connected to the source terminal. By changing the substrate voltage, we can alter the threshold voltage of the device because the gate charge must mirror Q_d before an inversion layer is formed, which is called the **body effect**. It can be derived that

$$V_{TH} = V_{TH0} + \gamma\left(\sqrt{2\Phi_F + V_{SB}} - \sqrt{2\Phi_F}\right) \tag{2.18}$$

where

(1) $\gamma = \frac{\sqrt{2q_e\varepsilon_{si}N_{sub}}}{C_{ox}}$: the **body effect coefficient**, typically 0.4 $V^{1/2}$;

(2) $\Phi_F = \frac{kT}{q_e}\ln\frac{N_{sub}}{n_i}$: the Fermi potential difference between the surface and bulk.

2.3.2 Channel-Length Modulation

I_D in the saturation region, given by (2.7) and (2.10), is actually a function of V_{DS}, which is called **channel-length modulation**. Writing the actual length L' as $L' = L - \Delta L$, i.e., $\frac{1}{L'} \approx \frac{1+\frac{\Delta L}{L}}{L}$, we have (in saturation)

$$I_D \approx \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.19}$$

where $\lambda \propto \frac{1}{L}$ is the **channel-length modulation coefficient**. λ represents the relative variation in length for a given increment in V_{DS}. Thus, for longer channels, λ is smaller. Note that there is no channel-length modulation in triode region.

To consider channel modulation, simply change $\frac{1}{L}$ to $\frac{1+\lambda V_{DS}}{L}$ to modify the previous equations for better accuracy.

Nanometer transistors suffer from various imperfections and markedly depart from square-law behavior. Shown below are the actual I-V characteristics of an NFET with $\frac{W}{L} = \frac{5\ \mu m}{40\ nm}$ for $V_{GS} = 0.3\ \text{V}, ..., 0.8\ \text{V}$. Also plotted are the characteristics of a square-law device of the same dimensions. Despite our best efforts to match the latter device to the former, we still observe significant differences.

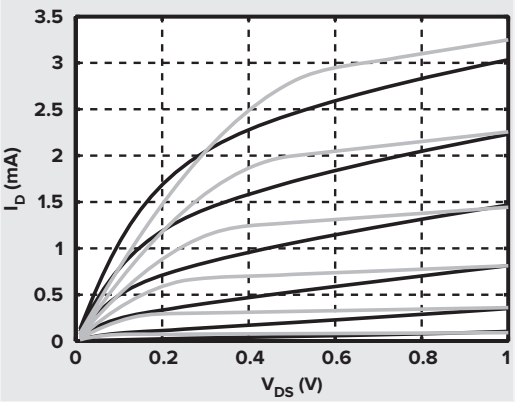


Figure 2.10: Actual I-V characteristics of an NFET and a square-law device.

2.3.3 Subthreshold Conduction

In reality, for $V_{GS} \approx V_{TH}$, a **week inversion layer** will still exist. Even for $V_{GS} < V_{TH}$, I_D is not zero, but exhibiting an expotional dependence on V_{GS}, called **subthreshold conduction**. Assuming V_{DS} is large enough ($V_{DS} > 100$ mV), the drain current is given by

$$I_D = I_0 \exp \frac{V_{GS}}{\xi V_T} = \alpha \frac{W}{L} \exp \frac{V_{GS}}{\xi V_T} \tag{2.20}$$

where

(1) $I_0 \propto \frac{W}{L}$: the drain current at $V_{GS} = V_{TH}$;
(2) $\xi > 1$: the noideality factor (typically 1.5);

As shown in Figure 2.11, we extrapolate the transfer characteristics I_D-V_{GS} on a logarithmic scale and consider their intercept voltage as the threshold voltage.

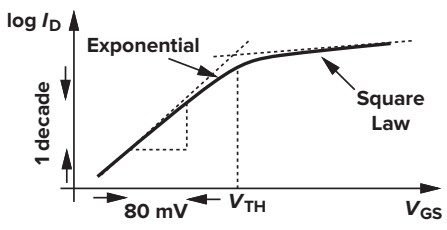


Figure 2.11: MOS subthreshold characteristics

When the corresponding transconductances g_m become equal for the same drain current, we say the transistor switch the inversion region. More specifically

$$\frac{I_D}{\xi V_T} = \frac{2I_D}{(V_{GS} - V_{TH})_{\text{switch}}} \tag{2.21}$$

$$(V_{GS} - V_{TH})_{\text{switch}} = 2\,\xi V_T \approx 80\ \text{mV} \tag{2.22}$$

We say the device operates in **week inversion** for $V_{GS} \leqslant (V_{TH} + 80\ \text{mV})$, and similarly in **strong inversion** for $V_{GS} > (V_{TH} + 80\ \text{mV})$. Equation (2.20) indicates that V_{GS} must decrease by roughly 80 mV for I_D to decrease by one decade (at room temperature), resulting a significant leak current (or power dissipation) for low threshold voltage devices.

To determine the operation region, for a given I_D, we need to obtain V_{GS} from both the suquare-law and the exponential-law and select the lower value:

$$V_{GS} = \min \left\{ \sqrt{\frac{2I_D}{\mu_n C_{ox} \frac{W}{L}}} + V_{TH},\ \xi V_T \ln \frac{I_D}{\alpha \frac{W}{L}} \right\} \tag{2.23}$$

2.4 MOS Device Models

2.4.1 MOS Device Layout

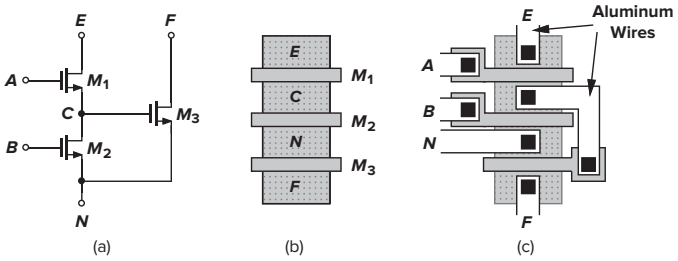


Figure 2.12: Layout of an NMOS device.
(a) bird's-eye view;
(b) top view (vertical view).

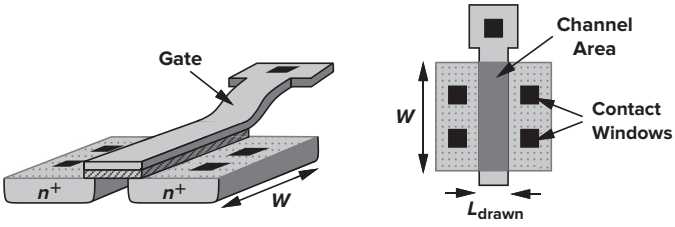


Figure 2.13

2.4.2 MOS Device Capacitances

To predict the high-frequency behavior, depicted in Figure 2.15, we expect that a capacitance exists between every two terminals of a MOSFET (C_{DS} is negligible). And these capacitances may depend on the bias conditions.

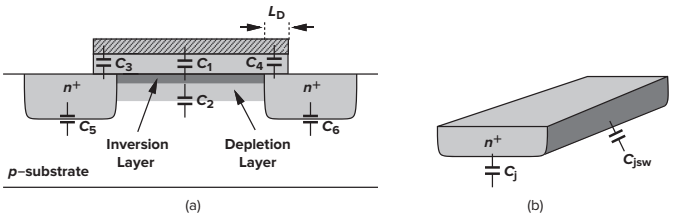


Figure 2.14: Decomposition of MOS capacitances
(a) MOS device capacitances
(b) S/D junction capacitance into bottom-plate and sidewall components

As shown in the figure above, we have:

(1) $C_1 = WLC_{ox}$: the oxide capacitance between the gate and the channel;

(2) $C_2 = WL\sqrt{\frac{q_e \varepsilon_{\text{si}} N_{\text{sub}}}{4\Phi_F}}$: the depletion capacitance between the channel and the substrate;

(3) $C_3 = C_4 = WC_{ov}$: the overlap capacitance, where C_{ov} is the overlap capacitance per unit width;

(4) $C_j = \frac{C_{j0}}{(1+\frac{V_R}{V_{\text{built-in}}})^m}$: the junction bottom-plate capacitance per unit area, where m typically in the range of 0.3 and 0.4;

(5) $C_{jsw} = \frac{C_{jsw0}}{(1+\frac{V_R}{V_{\text{built-in}}})^m}$: the junction sidewall capacitance per unit area;

(6) $C_5 = C_6 = S_{bp}C_j + S_{sw}C_{jsw}$: the junction capacitance

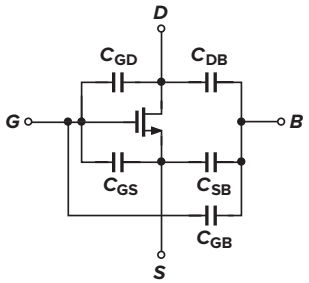


Figure 2.15: MOS capacitances

The terminal capacitances can be concluded as follows:

$$C_{DB} = C_{SB} = C_5 = S_{bp}C_j + S_{sw}C_{jsw}$$

$$C_{GB} = \begin{cases} C_1 \text{ series } C_2, & \text{off} \\ 0, & \text{else} \end{cases}$$

$$C_{GD} = \begin{cases} C_3 = WC_{ov}, & \text{else} \\ \frac{1}{2}C_1 + C_3 = \frac{1}{2}WLC_{ox} + WC_{ov}, & \text{deep triode} \end{cases}$$

$$C_{GS} = \begin{cases} C_3 = WC_{ov}, & \text{off} \\ \frac{2}{3}C_1 + C_3 = \frac{2}{3}WLC_{ox} + WC_{ov}, & \text{saturation} \\ \frac{1}{2}C_1 + C_3 = \frac{1}{2}WLC_{ox} + WC_{ov}, & \text{deep triode} \end{cases}$$

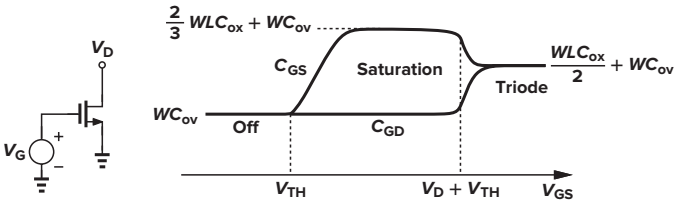


Figure 2.16: Variation of C_{GS} and C_{DS} versus V_{GS}

If the bulk terminal is tied to source, then C_{SB} is shorted, C_{DB} becomes C_{DS}, and C_{GB} is added to the original C_{GS}, yielding

$$C_{DS} = C_5 = S_{bp}C_j + S_{sw}C_{jsw}$$

$$C_{GD} = \begin{cases} C_3 = WC_{ov}, & \text{else} \\ \frac{1}{2}C_1 + C_3 = \frac{1}{2}WLC_{ox} + WC_{ov}, & \text{deep triode} \end{cases}$$

$$C_{GS} = \begin{cases} (C_1 \text{ series } C_2) + C_3, & \text{off} \\ \frac{2}{3}C_1 + C_3 = \frac{2}{3}WLC_{ox} + WC_{ov}, & \text{saturation} \\ \frac{1}{2}C_1 + C_3 = \frac{1}{2}WLC_{ox} + WC_{ov}, & \text{deep triode} \end{cases}$$

As shown in the figure below, new generations of CMOS technology incorporate the **FinFET** structure. The transistor carries current from S to D on the surfaces of the fin. The FinFET exhibits less channel-length modulation and subthreshold leakage by sacrificing contacts land and some other parameters.

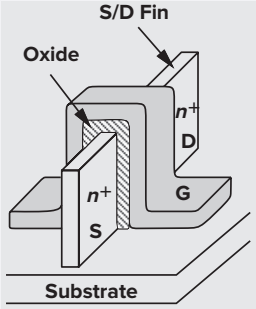


Figure 2.17: FinFET structure

2.4.3 MOS Small-Signal Model

$g_m = \frac{\partial I_D}{\partial V_{GS}}$ describe the small-signal behavior of MOSFETs when there is a perturbation in V_{GS}. Owing to channel-length modulation, I_D also varies with V_{DS}, and it can be modeled by a linear resistor r_O, given by

$$r_O = \frac{\partial V_{DS}}{\partial I_{DS}} = \begin{cases} \left[\mu_n C_{ox}\frac{W}{L}\left(V_{GS} - V_{TH} - V_D S\right)\right]^{-1}, & \text{triode} \\ \left[\frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left(V_{GS} - V_{TH}\right)^2 \cdot \lambda\right]^{-1}, & \text{saturation} \end{cases} \tag{2.24}$$