

CHAPTER 3 Device Layer

器件层面

3.1 Introduction 引言

The device layer of electrical design is where the bias conditions and physical parameters of all devices are determined from specified values of device performance metrics as depicted in Fig. 1.2. For resistors and capacitors, the relation between these two sets of variables is accurately modeled by simple analytical models. For MOSFET models, however, accuracy and simplicity are mutually exclusive because device operation is much more complicated. First-order analytical MOSFET models are good for gaining an insight into the path connecting the performance metrics to bias conditions and physical parameters but they are grossly inaccurate. Designers utilize them for qualitative purposes only. All quantitative tasks of design are performed with Spice simulations utilizing accurate but complex MOSFET models. However, simulation is a bottom-up tool operating in the opposite direction of device-design flow. It accepts physical parameters and bias conditions as input variables, and puts out electrical variables that represent performance metrics. Inevitably, therefore, design by simulation turns into an iterative process. In each iteration, a simulation is run with some estimates of the physical parameters and bias conditions, and the outcome is compared with the target values specified for performance metrics. This process is repeated in cycles until the differences between the simulated and targeted values of performance metrics are reduced to an acceptable level. Unfortunately, the multidimensional nature of the design space can turn this process into an almost perpetual sequence of trial and error unless (a) the initial values of physical parameters and bias conditions are selected intelligently, and (b) the values of these design variables are updated in the proper direction and amount in subsequent iterations. After briefly reviewing MOSFET basics in Section 3.2, we develop a simulation-based design methodology in Sections 3.3 to 3.6 exactly for this purpose. Based on a few generic design tools and involving only a few and simple calculations at most, it will quickly guide the designer from performance metrics to the optimum set of physical parameters and bias conditions.

All examples presented from this point forward are based on a mature 0.18- μm bulk complementary metal-oxide-semiconductor (CMOS) technology. In Chapter 4 the reader is expected to rework some of the examples and exercises in 90-nm bulk CMOS. We will treat these as twin-well technologies in all examples and exercises except for Example 4.27, where a triple-well technology will be assumed. Although advanced CMOS technologies offer core MOSFETs of a minimum feature size much smaller than these, the maximum voltage rating of such devices happens to be too restrictive for analog and

mixed-signal circuits. This is why analog applications are usually implemented with core devices in mature technologies or with optional high-voltage *input/output* (I/O) devices in advanced technologies. The feature size of I/O devices is kept larger than core devices in order to accommodate the higher maximum voltage rating. 0.18- μm is one such option available even in 14-nm CMOS technologies.

As mentioned above, MOSFETs are represented by an accurate model in Spice simulations. The *Berkeley Short-Channel IGFET Model* (BSIM) family is an industry standard in this respect.[1][2] BSIM3v3 and BSIM4 serve pre-0.1- μm and post-0.1- μm CMOS technology nodes, respectively. A third version was released in 2013 for post-40 nm nodes as BSIM6, which was renamed BSIM-BULK in 2017. Each version features hundreds of device parameters, among which only channel dimensions are under the control of an electrical designer. The rest are constants characterized by the manufacturer and made available to the designer as a part of the *process design kit* (PDK). The list of these constants is known as *model deck* in Spice terminology. Actually, a PDK contains multiple model decks describing several different versions of core and I/O devices for design flexibility. These versions differ mainly in the threshold voltage and are known as *threshold options*. Most typically, zero-threshold, low-threshold, and standard-threshold options are offered. Most manufacturers further diversify the model deck of each option by dividing the design space of device dimensions into subspaces, and assigning a slightly different deck to each. This practice of *binning* improves model accuracy. Still further, a deck is replicated for the worst-case and best-case values of parameters representing the extreme cases of variation the selected technology may exhibit among its runs. These are used in so-called *corner simulations* for verifying the robustness of the electrical design. All in all, a large number of model decks are usually present in a PDK. Fortunately, the simulation-based design methodology presented in this chapter is independent of the quantity or complexity of model decks. Based on this fact, we introduce the methodology with only two typical decks describing the standard-threshold NMOS and PMOS devices of a 0.18- μm CMOS technology but will apply it later in Chapter 4 also with low-threshold versions of these decks. The standard-threshold parameter decks are presented in Table 3.1 in a format recognizable as a model declaration by Spice. Saved in a file `bsim.sp`, these decks can be included automatically in any Spice input file containing the line `.incl bsim.sp`. Most of the parameters included in these decks have been characterized on fabricated wafers and made public by the *MOSIS Service*. [3]

Relating open-loop metrics to opamp device metrics is a task undertaken in the circuit layer of electrical design, and will be studied in Chapter 4. However, its proper execution necessitates a firm understanding of the techniques used in MOSFET small-signal analysis. Taking the opportunity provided by the first five sections of the present chapter as a background, we present these techniques in Section 3.7. MOSFET coverage continues in Section 3.8 where we present a model for the inherent noise generated by this device and exemplify its use in determining the noise performance of an analog circuit. Section 3.9 is devoted to the switching performance of MOSFET as an external network component of discrete-time configurations. In Section 3.10, we discuss the design of integrated resistors which find use not only in external networks but also inside the opamp. Finally, in Section 3.11, we move into integrated capacitor structures, which are indispensable not only for constructing the external networks of discrete-time closed-loop configurations but also for stabilizing certain opamp topologies.

```

.MODEL CMOSN NMOS (
+TOX = 4.1E-9
+K1 = 0.5789116
+W0 = 1E-7
+DVT2W = 0
+U0 = 293.1687573
+VSAT = 1.676164E5
+B1 = 5E-6
+RDSW = 105.6133217
+WINT = 2.885735E-9
+DWG = 2.754317E-9
+CIT = 0
+ETA0 = 2.665034E-3
+PDIBLC1 = 0.3258185
+PSCBE1 = 4.494778E10
+RSH = 7
+KT1 = -0.11
+UB1 = -7.61E-18
+WLN = 1
+LL = 0
+LWL = 0
+CGSO = 8.58E-10
+MJ = 0.3726161
+CJSWG = 3.3E-10
+PVTH0 = -5.105777E-3
+LKETA = 5.324922E-4
+PVSAT = 2E3
+NOIA = 1.2E+19
+EM = 4.1E7
)
*
.MODEL CMOSP PMOS (
+TOX = 4.1E-9
+K1 = 0.5722049
+W0 = 1E-6
+DVT2W = 0
+U0 = 109.4682454
+VSAT = 1.054892E5
+B1 = 1.446715E-6
+RDSW = 199.1594405
+WINT = 0
+DWG = -1.998034E-8
+CIT = 0
+ETA0 = 3.515392E-4
+PDIBLC1 = 3.026627E-3
+PSCBE1 = 7.999986E10
+RSH = 8.1
+KT1 = -0.11
+UB1 = -7.61E-18
+WLN = 1
+LL = 0
+LWL = 0
+CGSO = 7.82E-10
+MJ = 0.4192076
+CJSWG = 4.22E-10
+PVTH0 = 5.167913E-4
+LKETA = -3.648003E-3
+PVSAT = 50
+NOIA = 2.6E+20
+EM = 1.1E7
)
LEVEL = 49
XJ = 1E-7
K2 = 1.110723E-3
NLX = 2.037748E-7
DVT0 = 1.2953626
UA = -1.21942E-9
AO = 2
KETA = -0.0138552
PRWG = 0.5
LINT = 1.715622E-8
DWB = -3.690793E-9
CDSC = 2.4E-4
ETAB = 6.028975E-5
PDIBLC2 = 2.701992E-3
PSCBE2 = 3.672074E-8
MOBMOD = 1
KT1L = 0
UC1 = -5.6E-11
WW = 0
LLN = 1
CAPMOD = 2
CGBO = 1E-12
CJSW = 1.905901E-10
PBSWG = 0.8
PRDSW = -1.1011726
PU0 = -4.0206081
PETA0 = 1E-4
NOIB = 9.58E4
)
VERSION = 3.2
NCH = 2.3549E17
K3 = 1E-3
DVTOW = 0
DVT1 = 0.3421545
UB = 2.325738E-18
AGS = 0.4764546
A1 = 1.09168E-3
PRWB = -0.2
XL = 0
VOFF = -0.0948017
CDSCD = 0
DSUB = 0.0442223
PDIBLCB = -0.1
PVAG = 0.0122755
PRT = 0
KT2 = 0.022
AT = 3.3E4
WWN = 1
LW = 0
XPART = 0.5
CJ = 9.471097E-4
PBSW = 0.8
MJSWG = 0.1369758
PK2 = 2.247806E-3
PUA = -4.48232E-11
PKETA = -2.090695E-3
NOIC = 1E-14
TNOM = 27
VTH0 = 0.3694303
K3B = 0.0297124
DVT1W = 0
DVT2 = 0.0395588
UC = 7.061289E-11
B0 = 1.617101E-7
A2 = 0.3303025
WR = 1
XW = -1E-8
NFACTOR = 2.1860065
CDSCB = 0
PCLM = 1.746064
DROUT = 0.9787232
DELTA = 0.01
UTE = -1.5
UA1 = 4.31E-9
WL = 0
WWL = 0
LWN = 1
CGDO = 8.58E-10
PB = 0.8
MJSW = 0.1369758
CF = 0
WKETA = -5.071892E-3
PUB = 5.018589E-24
NOIMOD = 2
EF = 1.02

```

TABLE 3.1 BSIM3v3.2 Model Deck for Standard-Threshold Devices

3.2 MOSFET Basics MOS晶体管基础

In this section, we first present a brief description of CMOS structure, and subsequently identify MOSFET terminals, electrical ports, and port variables. The section concludes with the definition of MOSFET performance metrics and design variables.

3.2.1 Structure and Electrical Ports 器件结构和电学端口

CMOS Structure

Cross-sectional and top views of PMOS and NMOS transistor structures in a typical twin-well bulk CMOS technology are shown in Fig. 3.1. At the start of fabrication, the substructure consists of the p-substrate only while no superstructure is yet in place. The structural features first built into the substructure are the silicon dioxide-filled *shallow trench isolation* regions marked STI in the cross-sectional view. These trenches provide lateral isolation between neighboring MOSFETs. The area they collectively occupy on chip surface is called *field*, which complements the so-called *active* area occupied by MOSFETs and contact regions. Next in fabrication, *n-well* regions of PMOS devices and *p-well* regions of NMOS devices are formed by selectively doping the surface of the p-substrate. N-well is needed for providing isolation between the p-type source and drain of a PMOS device.

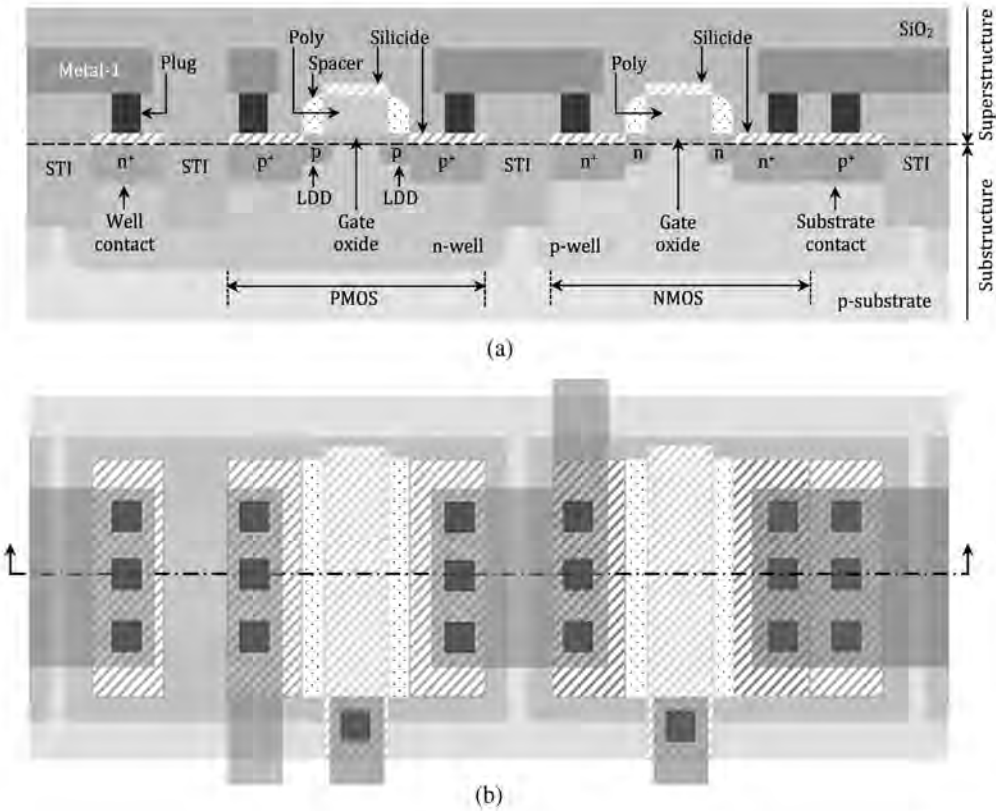


FIGURE 3.1 Twin-well bulk CMOS structure. (a) Cross-sectional view. (b) Top view and line of sectioning.

P-well is not intended for isolation because the n-type source and drain regions of an NMOS device could be isolated by the p-substrate alone, but the presence of a p-well makes it possible to optimize the performance of the NMOS device independently from the PMOS. Next in line of processing is the growth of *gate oxide* as the first structural feature of the superstructure. This is followed by the deposition and patterning of the polycrystalline silicon gate material, which is called *poly* in short. Source and drain regions of both devices are self-aligned with respect to poly gates in a two-step doping process. First is the formation of extensions called lightly doped drain (LDD). These are marked *p* and *n* on the cross-sectional view. After masking with a *sidewall spacer*, a second doping creates the relatively deeper and more conductive body of the source and drain regions, which are marked with p^+ and n^+ . This doping process is used also for raising the dopant concentration of n-well and p-well regions where metal contacts are to be made at a later stage of fabrication. Without doping enhancement, these contacts would exhibit a rectifying property instead of becoming ohmic. These regions are identified as *well contact* and *substrate contact* on the cross-sectional view. As exemplified on the NMOS structure, a contact region may be abutting a source region provided that both are to operate at the same potential. Otherwise, they are laid out to be disjoint as exemplified in the PMOS structure. The next step of fabrication is the conversion of the surface of the source, drain, and contact regions and poly into *silicide* after capping typically with titanium and heating in an inert environment. The conductance of these layers is significantly increased by this process. Next, a layer of silicon dioxide is deposited onto the surface, and holes are etched through where metal contacts are to be made to the substructure and poly. In order to create a contact, each of these holes is filled with a *plug* made typically of tungsten. This is followed by the deposition and patterning of interconnection metal, typically copper or aluminum. Notice the presence of a single layer of metal, marked *metal-1*, in Fig. 3.1. In actual fact, several layers of metal are available in CMOS technologies. These layers are separated by additional silicon dioxide layers, and are interconnected through holes filled with plugs and called *vias*.

Channel Dimensions

Simplified versions of cross-sectional and top views of NMOS and PMOS devices are presented in Fig. 3.2 for the purpose of defining the reference directions of electrical variables. Also shown are the symbols representing these devices in analog circuit schematics. Note that the current of a MOSFET is controlled inside the *channel*, which is how we name the region in between source and drain. The channel dimension along the direction of current flow between source and drain is called *effective length*, and is denoted by L on the top view of the NMOS device in Fig. 3.2. The channel dimension that is orthogonal to the direction of current flow is called *effective width*, and is denoted by W . As will be explained in Subsection 3.3.1, in all operation modes, MOSFET current is proportional to the so-called *aspect ratio* W/L .

W and L represent the final dimensions of a fabricated channel. They generally differ from the so-called *drawn dimensions* W_{DR} and L_{DR} which are quantified in electrical design and registered on the artwork in physical design. The difference, called *process bias*, is due to dimensional offsetting effects experienced during and after pattern transfer from artwork to silicon in fabrication. The relations between effective and drawn dimensions are described by

$$W = W_{DR} + \alpha W - 2 \times \overline{WINT}, \quad (3.1)$$

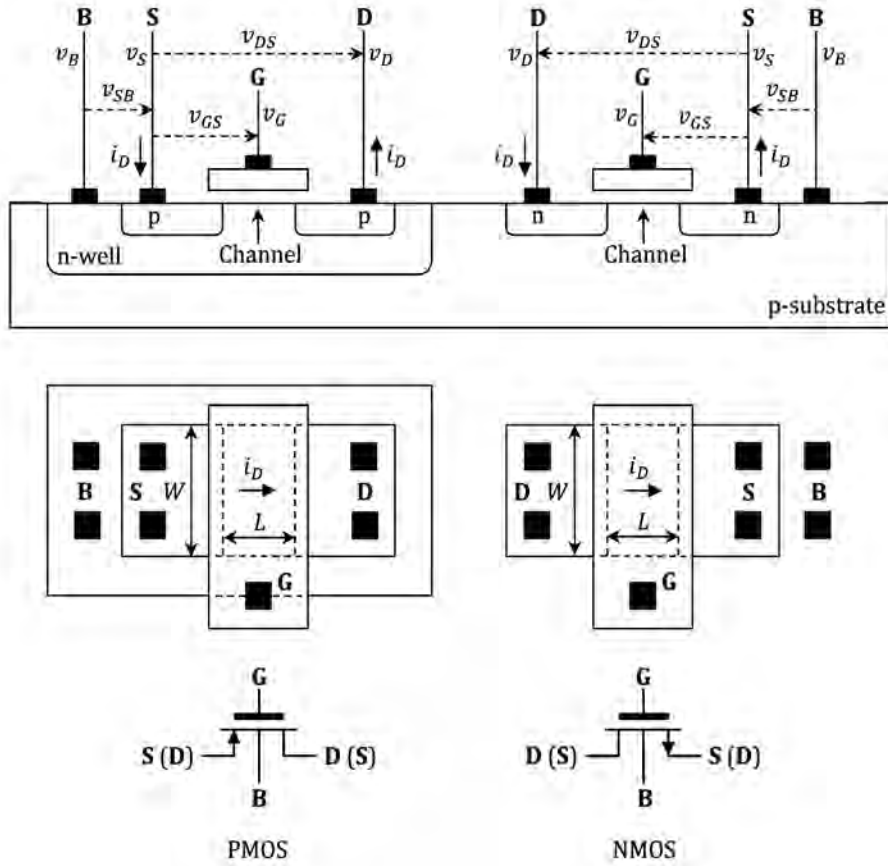


FIGURE 3.2 Top: Simplified cross-sectional view and electrical port variables of NMOS and PMOS structures. Middle: Top view of the structures. Bottom: Symbols representing these devices in circuit schematics.

and

$$L = L_{DR} + XL - 2 \times LINT, \quad (3.2)$$

where XW (m) and XL (m) represent the offsets created during pattern transfer, whereas $WINT$ (m) and $LINT$ (m) are those created by subsequent processing. All four are BSIM parameters whose values for our exemplar technology can be found in Table 3.1. It is also very important to note and remember that the instance-parameter symbols W and L used in MOSFET element lines of Spice input files represent not effective but drawn dimensions.

Terminals

A MOSFET has four terminals for electrical access. Namely, drain (D), gate (G), source (S), and bulk (B), whose voltages with respect to ground are denoted with v_D , v_G , v_S , and v_B , respectively. Bulk terminal is the p-substrate in an NMOS and n-well in a PMOS.

These four terminals constitute three independent electrical ports for each device. For modeling purposes, it is most appropriate to define gate source, drain source, and source bulk as independent ports.

Gate-Source Port

The voltage $v_{GS} \doteq v_G - v_S$ of this port is the main determinant of the concentration of current carriers hence conductivity inside the channel. It controls the concentration by imposing a transversal electric field across the gate oxide between gate and channel. In an NMOS, whose carriers are negatively-charged electrons, an increasingly positive v_{GS} enhances the conductivity by attracting a larger concentration of carriers. In a PMOS, positively-charged holes are current carriers. Therefore, an increasingly negative v_{GS} is needed for establishing a larger carrier concentration and thus enhancing the conductivity. The insulating property of the gate oxide prevents this port from conducting any dc steady-state current itself.

Drain-Source Port

The voltage $v_{DS} \doteq v_D - v_S$ of this port is responsible for driving the carriers from source to drain inside the channel, and thus generating the so-called *drain current* i_D . Since carriers are negatively charged in an NMOS, and positively charged in a PMOS, the polarity of the port voltage must satisfy the condition

$$\begin{aligned} v_{DS} &> 0 \text{ in NMOS,} \\ v_{DS} &< 0 \text{ in PMOS} \end{aligned} \tag{3.3}$$

in order to drive them from source to drain. By the same token, the resulting drain current is directed from drain to source in an NMOS, and from source to drain in a PMOS. Also note that $v_{DS} = 0$ implies $i_D = 0$ because carrier motion between source and drain is impossible without a driving force.

The source and drain terminals of a MOSFET are structurally undistinguishable. We use (3.3) as the basis of drain/source designation when we inspect a circuit schematic. Between the two terminals of the drain-source port, the one at the higher potential is designated as drain in NMOS and source in PMOS. If both are at the same potential ($v_{DS} = 0$), it does not matter which way we designate them. Majority of MOSFETs operate with one of these terminals continuously staying at a higher potential than the other. The drain/source designation of such devices never changes during operation. In some MOSFETs, however, the polarity of the potential difference between these two terminals is transposed during operation. An important implication of (3.3) for such a MOSFET is that source/drain designation must also be transposed if and when this happens. For this reason, the arrowhead appearing in analog MOSFET symbols must be interpreted with caution. It must be recognized solely as an indicator of whether the device is an NMOS (outward arrow) or a PMOS (inward arrow), and not necessarily as an indicator of the source terminal.¹

¹ The arrowhead of the analog MOSFET symbol is inherited from the BJT symbol, where it identifies not only the type of device as *pnp* or *npn* but also the emitter terminal.

Source-Bulk Port

Carrier population and therefore conductivity inside the channel can be controlled also by the source-bulk port voltage $v_{SB} \triangleq v_S - v_B$. For a fixed v_{GS} , conductivity decreases as v_{SB} is made more positive in NMOS or more negative in PMOS. However, the efficacy of this so-called *body effect* is about an order of magnitude weaker than that of the gate-source port, which is why source-bulk port is not intentionally utilized for controlling i_D in mainstream applications. Also note that this port may conduct a significant current of its own if the source-bulk junction is forward biased. Such a current is undesirable because it would flow outside the channel beyond the control of the gate terminal. This is why in mainstream applications the polarity of v_{SB} is ensured by design to comply with

$$\begin{aligned} v_{SB} &\geq 0 \text{ in NMOS,} \\ v_{SB} &\leq 0 \text{ in PMOS.} \end{aligned} \tag{3.4}$$

Note that the condition set by (3.4) prevents also the drain-bulk junction from being forward biased because drain potential cannot be lower than source potential in an NMOS or higher than source potential in a PMOS as implied by (3.3).

The only notable exception where source-bulk port is intentionally utilized as a control port is the so-called *bulk-driven MOSFET* application, which occasionally finds use in low-voltage design. This application also violates the condition (3.4).

Bulk Connection

Drain, gate, and source are the primary terminals by which a MOSFET is interconnected with other devices. Bulk is secondary in all mainstream applications but still it must be tied to an appropriate potential satisfying the condition (3.4). For a discussion of the options available in this regard, consider the general case of a split-supply circuit operating between a positive power-supply rail of voltage V_{DD} and a negative power-supply rail of voltage V_{SS} . Four possible ways in which a PMOS device can be interconnected with the rest of the circuit are represented in Fig. 3.3(a) with M1, M2, M3, and M4, all of which have a common bulk tied to V_{DD} . Condition (3.4) is satisfied by all these devices because no source terminal can assume any potential higher than V_{DD} . The source-bulk port of M1 remains unbiased ($v_{SB} = 0$), while those of the other three are reverse biased ($v_{SB} < 0$). This arrangement also yields a very compact layout because we do not need a separate n-well for each device; a single n-well tied to V_{DD} can serve as a common bulk for all.

As explained above, increasing the reverse bias of the source-bulk port leads to a lower channel conductivity through body effect even if v_{GS} is kept constant. In those applications where this is intolerable, the bulk connection of M2, M3, and M4 can be arranged as shown in Fig. 3.3(b). In M2 and M3, upper terminals function permanently as a source because their potential cannot be less than those of the respective lower terminals. By placing each of these two devices in an individual well and short-circuiting the well to the upper terminal, we establish $v_{SB} = 0$ for both. Same can be done for M4 but only if its source/drain designation is invariant during operation. The dashed connection shown in Fig. 3.3(b) exemplifies such a case. It is worth noting that $v_{SB} = 0$ can be maintained in a PMOS even in the case of a changing source/drain designation but it necessitates a specific circuit identifying the source and connecting it to the well in real time.[4] However, the overhead of the additional circuit is justifiable only in very exceptional cases.

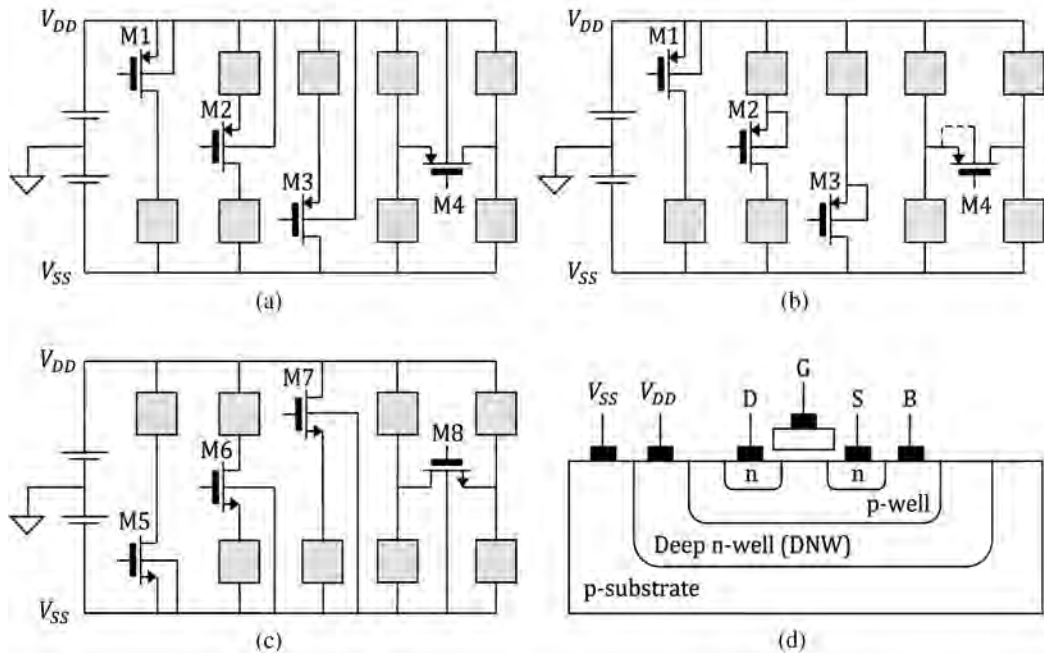


FIGURE 3.3 (a) All PMOS devices sharing a common bulk at V_{DD} . (b) Possibility of individual bulk-source connection in PMOS devices M1, M2, and M3. M4 can have an individual bulk-source connection if source designation does not change during operation. Otherwise, bulk is tied to V_{DD} . (c) All NMOS devices inevitably sharing a common bulk at V_{SS} . (d) A deep n-well NMOS structure in triple-well CMOS technology.

Next, consider Fig. 3.3(c), where all possible ways in which an NMOS device is interconnected with the rest of the circuit. Notice that the source of M5 is tied to the lowest potential on the chip. The only way we can satisfy (3.4) for this device is to connect the substrate also to V_{SS} . With this bulk connection, M5 will operate with $v_{SB} = 0$ but, since substrate is common to all NMOS devices, M6, M7, and M8 will operate with $v_{SB} > 0$. Condition (3.4) is therefore satisfied by all devices. Unlike the case of PMOS devices, however, it is impossible in a twin-well CMOS structure to establish $v_{SB} = 0$ for an NMOS device whose source potential is above V_{SS} as exemplified with M6, M7, or M8. This is because substrate is an inseparable common bulk for all NMOS devices in this technology; it cannot be individualized for a specific NMOS device. This limitation on NMOS devices is removed in the *triple-well* version of the CMOS technology, which offers a second type of NMOS device whose simplified cross-sectional view is shown in Fig. 3.3(d). Notice that the device is built inside a p-well, which, in turn, is built inside a deep n-well (DNW). By connecting DNW to V_{DD} , we keep it reverse biased with respect to the p-substrate, which stays at V_{SS} , and thus prevent forward biasing of the junction between the two. The isolated p-well can now be short-circuited to the source. Since their common potential never exceeds V_{DD} , the junction between p-well and DNW is never forward biased either. $V_{SB} = 0$ is therefore established for the NMOS device regardless of its source potential.

Example 3.1 Consider M8 of Fig. 3.3(c). How would you designate source and drain in the following cases of left-terminal potential V_L and right-terminal potential V_R ?

- $V_L = 1.2 \text{ V}$, $V_R = 0.8 \text{ V}$: Left terminal is drain.
- $V_L = -0.9 \text{ V}$, $V_R = -0.5 \text{ V}$: Right terminal is drain.
- $V_L = -0.3 \text{ V}$, $V_R = -0.7 \text{ V}$: Left terminal is drain.
- $V_L = 1.5 \text{ V}$, $V_R = 1.5 \text{ V}$: Indistinguishable. Designate arbitrarily.

Repeat the example for M4 of Fig. 3.3(a) and the following cases of terminal potentials:

- $V_L = 0.3 \text{ V}$, $V_R = 1.1 \text{ V}$: Left terminal is drain.
- $V_L = -1.3 \text{ V}$, $V_R = -0.5 \text{ V}$: Left terminal is drain.
- $V_L = 0.9 \text{ V}$, $V_R = 0.9 \text{ V}$: Indistinguishable. Designate arbitrarily.
- $V_L = 0.4 \text{ V}$, $V_R = -1.2 \text{ V}$: Right terminal is drain. ▲

Representation in Spice

The following element line is an example of how we generally declare a MOSFET instance in a Spice input file.

```
m18 3 12 7 8 cmosn w=2.5u l=0.7u ad=1.38p pd=6.1u nrd=0.08
+ as=1.38p ps=6.1u nrs=0.08 m=1
```

The line begins with the name “m18” of the instance where m identifies a MOSFET. The four subsequent characters “3 12 7 8” are the names of circuit nodes where the four terminals of the device are connected to. The second and fourth of these nodes are those of the gate and bulk, respectively. The drain and source terminals connect the first and third nodes but we are not concerned with which is which. In each simulation, Spice itself makes the correct assignment by comparing the voltages of these two nodes. What comes next in the element line is the model name of the device. The present example “cmosn” defines the NMOS device model of the deck presented in Table 3.1. The rest of the declarations in the element line are optional and not subject to any particular order. However, at least the drawn width “w” (m) and drawn length “l” (m) of the channel are necessary for any meaningful simulation. If not declared, Spice will assume a default value of 1 m for these parameters. “ad” (m^2) and “pd” (m) are, respectively, the area and perimeter of the source/drain region connecting the first of the four nodes. “nrd” is the number of squares traversed by device current inside this very same region. For understanding how we estimate these three instance parameters, consider the source/drain region tied to node 3 in the example layout of M18 shown in Fig. 3.4, and note

$$ad = W_{DR}L_1 = 2.5(\mu\text{m}) \times 0.55(\mu\text{m}) = 1.38 \times 10^{-12} \text{ m}^2,$$

$$pd = 2W_{DR} + 2L_1 = 2 \times 2.5(\mu\text{m}) + 2 \times 0.55(\mu\text{m}) = 6.1 \times 10^{-6} \text{ m},$$

and

$$nrd = L_{CG}/W_{DR} = 0.2(\mu\text{m})/2.5(\mu\text{m}) = 0.08.$$

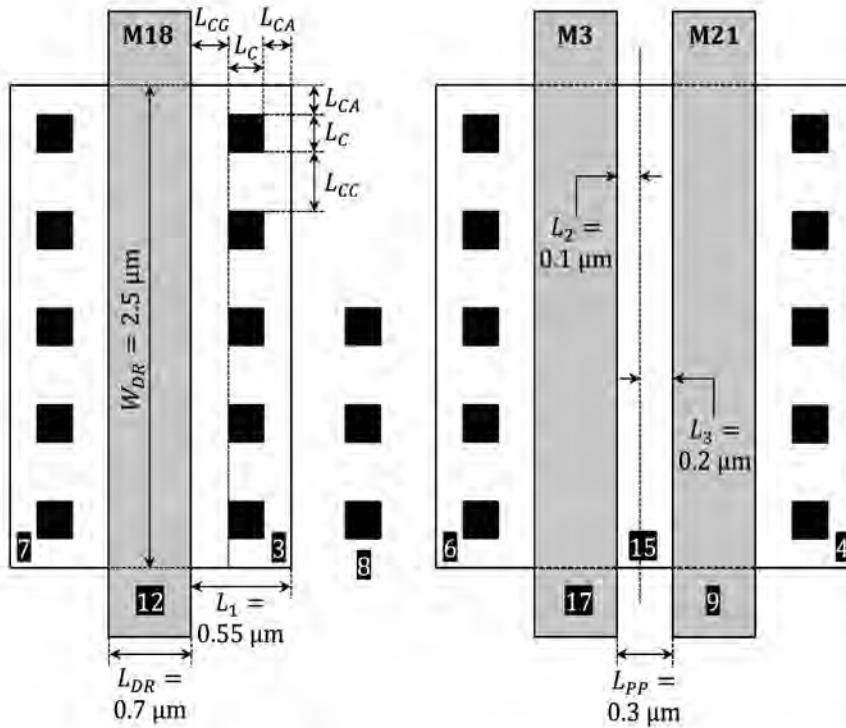


FIGURE 3.4 A layout example of three MOSFETs illustrating the design rules listed in Table 3.2. Gray-shaded features are polysilicon gates.

Notice the calculation of nrd by dividing the distance L_{CG} between the gate and inner side of contacts by the width W_{DR} of the region. Had the contacts been placed, say, one W_{DR} away from the gate, device current would traverse just 1 square in this region. In the present example, the division yields much lesser than 1 square because the contacts are pretty close to the gate. Returning to the exemplar element line, also notice the optional parameters as , ps , and nrs . These parameters represent the area, perimeter, and number of squares of the other source/drain region connecting the third of the four nodes. Since exemplar device M18 happens to be symmetrical in source/drain layout, the values of these parameters for node 7 are identical to their counterparts declared for node 3.

Spice needs the optional instance parameters ad , pd , as , and ps for taking into account certain parasitic effects associated with the source/drain regions. Junction capacitances, which are of foremost importance in transient, ac, and noise simulations, are among such parasitics. It is therefore essential in these types of simulation to declare the area and perimeter for any source/drain region whose voltage with respect to device bulk is not a constant. If not declared, Spice assumes a zero for each of these four parameters.

The optional instance parameters nrd and nrs enable Spice to determine the parasitic series resistances of source/drain regions by multiplying each with the model parameter

Number	Dimension	Symbol	Value
1	Side length of a contact	L_C	0.20- μm exact
2	Contact-to-contact spacing	L_{CC}	0.30- μm minimum
3	Contact-to-gate spacing	L_{CG}	0.20- μm minimum
4	Active overlap of contact	L_{CA}	0.15- μm minimum
5	Poly-to-poly spacing	L_{PP}	0.30- μm minimum

TABLE 3.2 Physical Design Rules for Source/Drain Layout

$r_{sh} (\Omega)$.² However, these resistances are of significance only in some exceptional cases. A device that is expected to have an unusually large contact-to-gate spacing and a large current is the most typical case. These parameters are needed also for accurately determining the very-high-frequency behavior of MOSFET input impedances as extracted with ac simulation. Also note that Spice assumes a default value of 1 for an undeclared nrd or nrs .

Since the actual values of ad , pd , nrd , as , ps , and nrs are finalized only after the device has been laid out in physical design, simulations conducted during electrical design should be based on the estimates of these instance parameters. For an accurate estimation, we need to be conversant with the physical-design rules pertinent to source/drain layout. Listed in Table 3.2 are such rules to be followed throughout this book. The first four of these rules are illustrated in Fig. 3.4 on device M18. Notice that the side length $L_1 = 0.55 \mu\text{m}$ used above for estimating ad and pd is the minimum permitted by rules 1, 3, and 4 for a source/drain region containing contacts. The length $L_{CG} = 0.2 \mu\text{m}$ of the parasitic resistance used for estimating nrd is also the minimum permitted by rule 3.

Each contact used for accessing a source/drain region has approximately 10Ω resistance in a typical 0.18- μm CMOS technology. Paralleling as many contacts as possible along the width W_{DR} not only helps minimize the equivalent resistance, but also provides redundancy against possible contact failure. Rule 2 is added to Table 3.2 for enabling the reader to determine the maximum number of contacts that can be accommodated by a source/drain region of a given W_{DR} .

Finally, notice rule 5 for the spacing between two parallel polysilicon lines. For an explanation of the utility of this rule, consider the layout of the two series-connected devices M3 and M21 in Fig. 3.4. These devices share a common source/drain region at node 15 without any metal interconnection, hence, without any contact. Since the polysilicon gates at nodes 17 and 9 can be laid out as close as permitted by rule 5 of Table 3.2, their common source/drain region can be as narrow as $L_{PP} = 0.30 \mu\text{m}$. This helps reduce the parasitic capacitance of this node nearly by half in comparison with a contacted source/drain region of the same W_{DR} , and is therefore practiced in physical design wherever two or more devices of the same conductivity type are connected in series without any interconnection branching out from their common nodes. Cascoded devices, as will be presented in Chapter 4, represent the most common case where this practice is followed in physical design. One outstanding question is how a common source/drain

²The reader is referred to Table 3.1 for the NMOS and PMOS values of r_{sh} in our exemplar technology.

region is to be partitioned between two devices in a Spice input file. The answer is that any fictitious line drawn widthwise across the region will do the partitioning. Although the positioning of the line is arbitrary, it is important to avoid assigning all of the region to one of the devices and none to the other because assigning a perimeter lesser than W_{DR} to any of the devices will result in an erroneous calculation of parasitic capacitance. In the example shown in Fig. 3.4, the line is drawn at a distance $L_2 = 0.1 \mu\text{m}$ from the gate of M3, and, therefore, at a distance $L_3 = 0.2 \mu\text{m}$ from the gate of M21. This leads to the element lines

```
m3 15 17 6 8 cmosn w=2.5u l=0.7u ad=0.25p pd=2.7u nrd=0.04
+ as=1.38p ps=6.1u nrs=0.08
m21 4 9 15 8 cmosn w=2.5u l=0.7u ad=1.38p pd=6.1u nrd=0.08
+ as=0.5p ps=2.9u nrs=0.08
```

where the dimensional parameters of the common-node 15 are estimated from $ad = W_{DR}L_2$, $pd = W_{DR} + 2L_2$, and $nrd = L_2/W_{DR}$ for M3, and from $as = W_{DR}L_3$, $ps = W_{DR} + 2L_3$, and $nrs = L_3/W_{DR}$ for M21. Note that the sum of ad of M3 and as of M21 equals the actual area of this common region. Similarly, the sum of pd of M3 and ps of M21 equals the actual perimeter. Same is true also for the sum of nrd of M3 and nrs of M21.

3.2.2 Performance Metrics and Design Variables 性能指标和设计参量

Inside opamps, MOSFETs usually function as transconductors or current sources/sinks. Device metrics associated with these two types of functionality are defined in this section. MOSFETs find use also as external-network switches in discrete-time configurations. The unique performance metrics of this type of application are discussed in Section 3.9. Occasionally, MOSFETs operate also as passive components in the form of a resistor or a capacitor. Performance metrics involved in these two types of operation are presented in Subsections 3.10.3 and 3.11.2, respectively.

A MOSFET operating as a transconductor converts a variation in the gate-source port voltage v_{GS} around a bias value V_{GS} into a variation in the drain-source port current i_D around a bias value I_D . This conversion is characterized with a performance metric called *transconductance* or, more specifically, *gate transconductance*, and defined by

$$g_m \triangleq \begin{cases} \left. \frac{\partial i_D}{\partial v_{GS}} \right|_{v_{GS}=V_{GS}} & \text{for NMOS,} \\ - \left. \frac{\partial i_D}{\partial v_{GS}} \right|_{v_{GS}=V_{GS}} & \text{for PMOS,} \end{cases} \quad (3.5)$$

where the sign convention yields a positive value for both types of MOSFET. Formally, g_m is measured in units of Siemens or A/V but the unit of preference in practice is $\mu\text{A/V}$ or mA/V .

In some transconductor applications, v_{SB} may also vary around a bias value V_{SB} , and cause a variation in i_D around a bias value I_D . The performance metric we use for characterizing this effect is called *bulk transconductance*. It is defined by

$$g_{mb} \triangleq \begin{cases} - \left. \frac{\partial i_D}{\partial v_{SB}} \right|_{v_{SB}=V_{SB}} & \text{for NMOS,} \\ \left. \frac{\partial i_D}{\partial v_{SB}} \right|_{v_{SB}=V_{SB}} & \text{for PMOS.} \end{cases} \quad (3.6)$$

The sign convention of (3.6) makes g_{mb} a positive-valued parameter for both types of MOSFET. g_{mb} is also measured in units of $\mu\text{A}/\text{V}$ or mA/V .

A MOSFET operating as a current source or a current sink is expected to keep its drain-source port current i_D constant at a nominal value I_D independently of the port voltage v_{DS} . I_D is the primary performance metric of this application. Its value is set by driving the gate-source port with a dc voltage V_{GS} . The source-bulk port is usually short-circuited but in rare cases it may hold a dc voltage V_{SB} . In practice, i_D cannot be made absolutely independent of v_{DS} . The extent of its dependence is characterized with two performance metrics. For a description of these two, the reader is referred to Fig. 3.5, where typical NMOS and PMOS *output characteristics* relating i_D to v_{DS} are shown. These characteristics indicate a reasonably independent current only for $v_{DS} > V_{DSsat}$ in NMOS and for $v_{DS} < V_{DSsat}$ in PMOS. The so-called *drain-source saturation voltage* V_{DSsat} obviously limits the domain in which a MOSFET can operate as a current source. It is also evident that the current supplied by a MOSFET is not perfectly independent of v_{DS} even if the limit set by V_{DSsat} is not breached. The rate by which it varies with v_{DS} is characterized with a second performance metric called *output conductance*, defined by

$$g_{ds} \doteq \begin{cases} \left. \frac{\partial i_D}{\partial v_{DS}} \right|_{v_{DS}=V_{DS}} & \text{for NMOS,} \\ - \left. \frac{\partial i_D}{\partial v_{DS}} \right|_{v_{DS}=V_{DS}} & \text{for PMOS,} \end{cases} \quad (3.7)$$

and measured in units of $\mu\text{A}/\text{V}$ or mA/V . The nonlinearity of the output characteristics makes g_{ds} a function of v_{DS} . This is why the definitions made in (3.7) are for a specific drain-source voltage V_{DS} . In most design scenarios, this will be the bias voltage across the drain-source port. Also note that the inverse of the output conductance, defined by $r_{ds} \doteq 1/g_{ds}$ (Ω) and called *output resistance* in this book, finds even more widespread use for characterizing the quality of the current source/sink performance.

In summary, the target performance of a MOSFET is specified in the circuit layer of electrical design with a subset of the following metrics: (a) dc drain current I_D , (b) drain-source saturation voltage V_{DSsat} , (c) transconductance g_m , (d) output conductance g_{ds} or its inverse r_{ds} , and (e) bulk transconductance g_{mb} . There exist at most five design variables at the disposal of a designer to implement the specified values of these metrics in

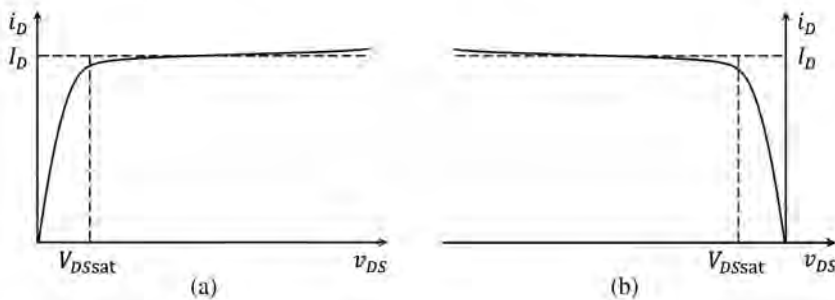


FIGURE 3.5 Typical variation of drain current i_D with drain-source voltage v_{DS} for a constant gate-source voltage V_{GS} in (a) NMOS and (b) PMOS.

the device layer of electrical design: (a) Drawn channel width W_{DR} , (b) drawn channel-length L_{DR} , (c) gate-source bias voltage V_{GS} , (d) drain-source bias voltage V_{DS} , and (e) source-bulk bias voltage V_{SB} . In many cases, V_{DS} and V_{SB} are already fixed by the topological properties of the opamp or constrained by the bias voltages of other devices. In some cases, even V_{GS} is imposed by another device. W_{DR} and L_{DR} however, are almost always available as independent design variables. Designing a MOSFET therefore turns into a process of sizing the device under constrained bias conditions. The relations connecting device performance metrics to design variables are the tools of this process. A detailed study of these relations and their use in NMOS device design are the subject of the following section. The PMOS version of the relations is presented in Section 3.4. The content of these two sections constitutes the basis of the device-design methodology presented subsequently in Section 3.6.

3.3 NMOS Design Relations and Tools NMOS晶体管设计方程和工具

The design relations presented in this section are derived from the dc relationship between i_D and its variables W , L , v_{GS} , v_{DS} , and v_{SB} . This relationship has a very complex makeup in BSIM models due to the inclusion of numerous secondary physical phenomena observed in devices of near-minimum channel dimensions. Fortunately, channel dimensions needed in analog applications are usually large enough to subdue most of these phenomena. Based on this premise, we present in Subsection 3.3.1 the so-called *long-channel models*, which offer a very simple analytical description of MOSFET operation by ignoring the secondary phenomena. However, these models are good enough only for drawing a qualitative picture of design relations because, as will be explained next in Subsection 3.3.1, certain additional assumptions built into their derivation are not fully justified in analog applications.

While long-channel models provide the designer with a qualitative insight, the quantitative tasks of design have to be based on simulation for accuracy. Starting with Subsection 3.3.2 and ending with Subsection 3.3.6, we will describe the way in which the performance metrics of NMOS devices can be related to design variables by simulation. Subsequently in Section 3.4, we will do the same for PMOS devices.

3.3.1 Long-Channel Models 长沟道模型

Long-channel models relate i_D to W , L , v_{GS} , v_{DS} , and v_{SB} for two asymptotic cases depending on the dominant physical mechanism of carrier flow inside the channel. In the case of *strong inversion*, carriers are assumed to flow purely by drifting under the influence of the lateral electric field created by a nonzero v_{DS} . This case prevails if carrier concentration at the source boundary of the channel is much larger than bulk-doping concentration. The second case, called *subthreshold*, is based on the assumption that the carriers move purely by diffusing thermally, which prevails when the carrier concentration at the same boundary is very low. Since these are two asymptotic cases of device operation, long-channel models are least accurate if the device operates in between the two. Since, unfortunately, this is where device performance is usually optimized in analog operation, we can utilize long-channel models only for qualitative purposes.

In this section, we present the analytical expressions of strong-inversion and sub-threshold models. Some of the parameters appearing in these expressions are available in the BSIM model deck presented in Table 3.1. Such parameters are printed in the typewriter font.

Strong-Inversion Model

The strong-inversion model divides the operation space of a MOSFET into three modes called *cutoff*, *saturation*, and *nonsaturation*.

Cutoff prevails if gate-source voltage is smaller than a *threshold voltage* parameter V_{TN} , that is,

$$v_{GS} - V_{TN} \leq 0 \text{ (condition of cutoff).} \quad (3.8)$$

Drain current is presumed to vanish in cutoff:

$$i_D = 0 \text{ (in cutoff).} \quad (3.9)$$

Threshold voltage is described by

$$V_{TN} = V_{TH0} + K1(\sqrt{2\phi_{FB} + v_{SB}} - \sqrt{2\phi_{FB}}), \quad (3.10)$$

which contains two BSIM parameters: (a) *Zero-bias threshold voltage* V_{TH0} (V) and (b) *body factor* $K1$ ($V^{1/2}$). ϕ_{FB} (V) is known as *Fermi potential*. It is described by

$$\phi_{FB} = \frac{kT}{q} \ln \frac{NCH}{n_i}, \quad (3.11)$$

where $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant, T (K) is absolute temperature, $q = 1.6 \times 10^{-19}$ C is the absolute value of electron charge, NCH (cm^{-3}) is a BSIM parameter called *substrate-doping concentration* (cm^{-3}), and n_i (cm^{-3}) is *intrinsic carrier concentration* whose value at room temperature ($T = 300$ K) equals $1.1 \times 10^{10} \text{ cm}^{-3}$. Note that kT/q (V) is known as *thermal voltage*; its equals 25.8 mV at room temperature.

As indicated by (3.10), V_{TN} is an increasing function of v_{SB} . This dependence is known as *body effect*. Also note that V_{TN} is reduced to V_{TH0} for $v_{SB} = 0$.

The device is in saturation if

$$v_{GS} - V_{TN} \geq 0 \text{ and } v_{DS} \geq V_{DSsat} \text{ (condition of saturation).} \quad (3.12)$$

Drain-source saturation voltage V_{DSsat} is related to v_{GS} through

$$V_{DSsat} \doteq \frac{v_{GS} - V_{TN}}{A_N}. \quad (3.13)$$

Although A_N is generally defined by

$$A_N \doteq 1 + \frac{K1}{2\sqrt{2\phi_{FB} + v_{SB}}}, \quad (3.14)$$

the following simplified but approximate description is more preferred for the purpose of evaluating V_{DSSat} :

$$A_N \approx 1 + \frac{K1}{2\sqrt{2\phi_{FB}}}, \quad (3.15)$$

In saturation, drain current is described by

$$i_D = \frac{W}{L} \frac{\mu_N C_{OX}}{2A_N} (v_{GS} - V_{TN})^2 \quad (\text{in saturation}), \quad (3.16)$$

where μ_N (m^2/Vs) represents *electron mobility*, which is a physical parameter relating the velocity of drifting carriers to lateral electric field. Although it is a decreasing function of v_{GS} , we treat it as a constant in hand calculations. A highly simplified form of its BSIM model is described by

$$\mu_N \approx \frac{U0}{1 + UA \frac{2 \times V_{TH0}}{TOX} + UB \left(\frac{2 \times V_{TH0}}{TOX} \right)^2}, \quad (3.17)$$

The units used for UA and UB in BSIM model decks are m/V and $(\text{m/V})^2$, respectively. U0, on the other hand, is cited in units of cm^2/Vs , which must be converted into m^2/Vs before being used in (3.17). Also appearing in (3.16) is C_{OX} (F/m^2), which represents the *oxide capacitance per unit gate area*. It is related to *gate-oxide thickness* TOX (m) by

$$C_{OX} \doteq \frac{\epsilon_{OX}}{TOX}, \quad (3.18)$$

where $\epsilon_{OX} = 3.5 \times 10^{-11} \text{ F/m}$ is the *dielectric constant* of the gate oxide.

An NMOS device is in nonsaturation if

$$v_{GS} - V_{TN} \geq 0 \text{ and } v_{DS} \leq V_{DSSat} \quad (\text{condition of nonsaturation}), \quad (3.19)$$

and the current is described by

$$i_D = \frac{W}{L} \mu_N C_{OX} \left(v_{GS} - V_{TN} - \frac{A_N}{2} v_{DS} \right) v_{DS} \quad (\text{in nonsaturation}), \quad (3.20)$$

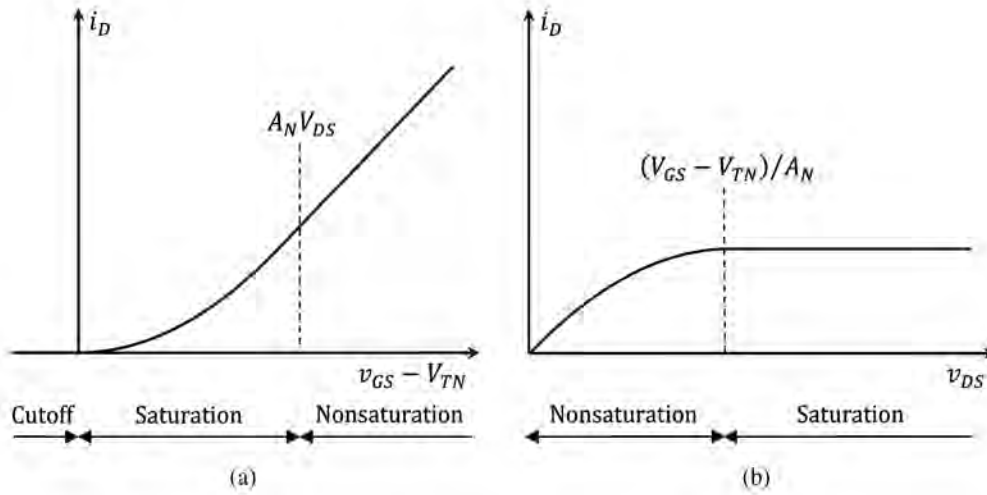


FIGURE 3.6 Strong-inversion model plots of NMOS drain-current versus (a) overdrive voltage, and (b) drain-source voltage.

Notice from (3.9), (3.16), and (3.20) the continuity of current-voltage characteristics at the boundaries $v_{GS} = V_{TN}$ and $v_{DS} = V_{DSsat}$ of the three operation modes. Also notice that v_{GS} never appears in any of the modeling equations without being subtracted by V_{TN} . The difference $v_{GS} - V_{TN}$ is indeed the most influential variable of MOSFET operation and deserves to have a specific name; we call it *overdrive voltage*.

Predictions of the long-channel model for strong inversion are illustrated in Fig. 3.6(a) with a plot of i_D versus overdrive voltage $v_{GS} - V_{TN}$ for a fixed drain-source voltage V_{DS} and a fixed source-bulk voltage V_{SB} . We observe (a) an absence of current in cutoff for negative values of the overdrive voltage $v_{GS} - V_{TN}$, (b) a quadratically increasing current in saturation for positive values of $v_{GS} - V_{TN}$ smaller than $A_N V_{DS}$, and (c) a linearly increasing current for positive values of $v_{GS} - V_{TN}$ greater than $A_N V_{DS}$. An alternative illustration of the same long-channel model predictions is presented in Fig. 3.6(b) with an output characteristic for a fixed positive overdrive voltage $V_{GS} - V_{TN}$ precluding cutoff and a fixed source-bulk voltage V_{SB} . Notice that i_D is predicted to be independent of v_{DS} only in the saturation mode prevailing for $v_{DS} \geq V_{DSsat} = (V_{GS} - V_{TN})/A_N$. In actual fact, it somewhat depends on v_{DS} even in saturation as explained in Subsection 3.2.2, but, compared with nonsaturation, the rate of dependence is much weaker. This is why saturation is essential in current-source and transconductor applications.

Subthreshold Model

Strong-inversion model is based on the assumption that drain current can flow only if the carrier concentration at the source boundary of the channel exceeds the doping concentration, which corresponds to the condition $v_{GS} - V_{TN} \geq 0$ in terms of NMOS overdrive voltage. This is why strong-inversion model predicts $i_D = 0$ unless this condition is met. In actual fact, any finite carrier concentration, however small it may be, yields a drain

current as long as $v_{DS} \neq 0$ but, of course, the smaller the concentration, the weaker the current. Therefore, a weak but finite current flows even under the cutoff condition of the strong-inversion model but the model ignores it. This is why model accuracy rapidly deteriorates as the overdrive voltage is lowered toward zero. The small but finite drain current flowing in cutoff is called subthreshold current, and is described approximately by

$$i_D = \frac{W}{L} \mu_N C_{OX} (A_N - 1) \left(\frac{kT}{q} \right)^2 \exp \left(\frac{q}{kT} \frac{v_{GS} - V_{TN}}{A_N} \right) \left[1 - \exp \left(-\frac{q}{kT} v_{DS} \right) \right], \quad (3.21)$$

without specifically distinguishing between saturation and nonsaturation modes. However, considering the current-source and transconductor applications of a MOSFET, we still need to identify the operation mode in which i_D can be considered reasonably independent of v_{DS} . Inspecting (3.21), we observe that such a mode prevails as long as v_{DS} is greater than about $2kT/q$. Therefore, we can define a saturation mode for subthreshold operation with

$$v_{GS} - V_{TN} < 0 \text{ and } v_{DS} \geq V_{DSsat} \text{ (condition of saturation in subthreshold)}, \quad (3.22)$$

where V_{DSsat} is given by

$$V_{DSsat} \approx 2 \frac{kT}{q}. \quad (3.23)$$

In saturation, (3.21) reduces to

$$i_D = \frac{W}{L} \mu_N C_{OX} (A_N - 1) \left(\frac{kT}{q} \right)^2 \exp \left(\frac{q}{kT} \frac{v_{GS} - V_{TN}}{A_N} \right) \quad (\text{in saturation}). \quad (3.24)$$

According to (3.21) or (3.24), i_D is an exponentially increasing function of the overdrive voltage $v_{GS} - V_{TN}$. However, this observation is an accurate representation of the actual subthreshold behavior only for $v_{GS} - V_{TN} \ll 0$ because (3.21) is derived under the assumption of a very small carrier population. Obviously, neither the strong-inversion model nor the subthreshold model can provide us with an accurate description of MOSFET behavior when the overdrive voltage is close to zero. This is unfortunate, because, as stated previously, the overall performance of a MOSFET in current-source and transconductor applications is maximized for small values of the overdrive voltage. We overcome this limitation by resorting to simulation for extracting i_D and other performance metrics wherever we need accuracy in device design. Incidentally, BSIM3 and BSIM4 also model strong inversion and subthreshold separately, and then, bridge the gap around zero-overdrive with mathematical extrapolation. Only in BSIM-BULK, a physically unified model is offered.

3.3.2 Threshold Voltage 阈值电压

The relation between the threshold voltage and design variables is one of the tools of device design. Recall from Subsection 3.2.2 that design variables include the drawn

dimensions W_{DR} , L_{DR} , and bias voltages V_{GS} , V_{DS} , and V_{SB} . According to (3.10) and (3.11), the sole design variable that affects V_{TN} among this set is V_{SB} . The rest of the entries of these equations are either physical constants or BSIM parameters, which, as explained in Section 3.1, depend only on temperature for a given threshold option and process corner. In practice, V_{TN} varies also with W_{DR} , L_{DR} , and V_{DS} but among these only the variation with L_{DR} is significant in analog applications. Therefore, a family of plots relating V_{TN} to V_{SB} for the typical range of channel length of analog applications can serve as a common design tool for all devices of the same threshold option and process corner. This design tool can be extracted accurately with a single Spice dc-sweep analysis. The only critical point in simulation is to set the width of the channel around what is typically encountered in practice. A good choice is $W_{DR} = 20W_{DR(min)}$, where $W_{DR(min)}$ is the *minimum drawn width* permitted by physical design rules. Selection of V_{GS} or V_{DS} is not critical; setting each to, say, one half of the maximum permissible port voltage V_{MAX} will do. Note $V_{MAX} = 1.8$ V for our exemplar technology.

Shown in Fig. 3.7 is the variation of V_{TN} with V_{SB} for eight different values of L_{DR} in our exemplar technology, in which, the minimum permissible dimensions are $W_{DR(min)} = 0.3$ μm and $L_{DR(min)} = 0.2$ μm . The following input file extracts and directly plots these curves:

```
filename: nmos1.sp (nmos vth)
.incl bsim.sp
m 2 1 3 0 cmosn w=6u l=0.2u
vds 2 3 0.9
```

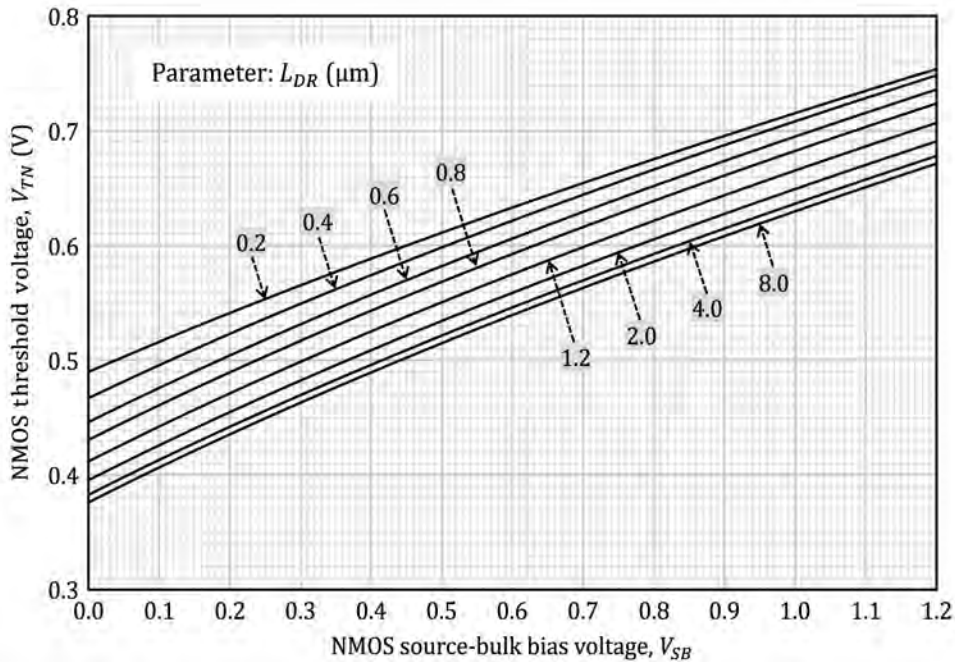


FIGURE 3.7 NMOS threshold voltage V_{TN} as a function of source-bulk bias voltage V_{SB} and drawn channel-length L_{DR} .

```

vgs 1 3 0.9
vsb 3 0 0
.save @m[vth]
.control
destroy all
foreach length 0.2u 0.4u 0.6u 0.8u 1.2u 2u 4u 8u
alter @m[l]=$length
dc vsb 0 1.2 20m
end
plot dc1.@m[vth] dc2.@m[vth] dc3.@m[vth] dc4.@m[vth]
+ dc5.@m[vth] dc6.@m[vth] dc7.@m[vth] dc8.@m[vth]
.endc
.end

```

Notice that V_{TN} is an increasing function of V_{SB} as predicted by (3.10) and a decreasing function of L_{DR} . In some technologies, however, it may be maximized at an L_{DR} larger than $L_{DR(min)}$. Also note that its value for $V_{SB} = 0$, known as *zero-bias threshold voltage*, is generally greater than the BSIM parameter $V_{TH0} = 0.37$ V which is what the long-channel model equation (3.10) predicts for $V_{SB} = 0$. As expected, however, the difference vanishes for large values of L_{DR} as the MOSFET becomes better qualified as a long-channel device.

3.3.3 Drain-Source Saturation Voltage 漏源饱和电压

Drain-source saturation voltage V_{DSsat} is among the device performance metrics defined in Subsection 3.2.2. Assuming the simplification offered by (3.15), the only design variable indicated by (3.13) for this metric in strong inversion is the overdrive voltage $V_{GS} - V_{TN}$. In subthreshold, (3.23) indicates a constant V_{DSsat} . Actually, V_{DSsat} also varies with L_{DR} and V_{SB} but these variations are secondary for the typical values of channel length and overdrive voltage involved in analog applications. A single plot of V_{DSsat} versus $V_{GS} - V_{TN}$ can therefore be used as a common design tool for all devices of a given set of threshold option, process corner, and temperature. We can extract the plot with a single simulation, and use it as a reasonably accurate design tool for any MOSFET of the same threshold option, same process corner, and same temperature just as we did in Subsection 3.3.2 for V_{TN} .

The simulation is run by setting $V_{SB} = 0$, and sweeping V_{GS} . As long as V_{DSsat} is plotted as a function of $V_{GS} - V_{TN}$, the plot will also be valid to a good approximation for those cases in which $V_{SB} \neq 0$. Since, in most design scenarios, overdrive voltage turns out to be less than one half of the maximum port voltage V_{MAX} , it usually suffices to set the upper limit of sweep to $V_{GS(max)} = V_{TN} + (V_{MAX}/2)$. The lower limit of sweep can be as low as $V_{GS(min)} = V_{TN}$ or even lower if a subthreshold operation is anticipated. Selection of V_{DS} is not critical; $V_{MAX}/2$ is a reasonable choice. As to drawn dimensions, $L_{DR} = 4L_{DR(min)}$, $W_{DR} = 20W_{DR(min)}$ are reasonable values. The input file `nmos2.sp` used for extracting V_{DSsat} for our exemplar MOSFET option is as follows:

```

filename: nmos2.sp (nmos vdsat, id, gm, gmbs)
.incl bsim.sp
m 2 1 3 0 cmosn w=6u l=0.8u
vds 2 3 0.9
vgs 1 3

```

```

vsb 3 0 0
.dc vgs 0.2 1.3 10m
.save @m[vth] @m[vdsat] @m[id] @m[gm] @m[gmb] v(1,3)
.end

```

Note that this file is also used in the following three subsections for developing design tools associated with metrics I_D , g_m , and g_{mb} . This is why these three metrics are also saved in addition to drain-source saturation voltage @m[vdsat] and gate-source voltage v(1,3). The command

```
plot @m[vdsat] vs v(1,3)-@m[vth]
```

will generate the plot shown in Fig. 3.8. Qualitatively, the variation is consistent with the prediction of (3.13) for large values of overdrive implying strong inversion and that of (3.23) for small values implying subthreshold. The plot smoothly connects the two in weak inversion where each model is most inaccurate.

As explained in Subsection 3.2.2, V_{DSsat} represents the lowest drain-source voltage for which an NMOS device can remain in saturation as required in current-source and transconductor applications. Therefore, a wide range of V_{DS} is possible in these two types of applications only if V_{DSsat} is small. We now know that this calls for a small overdrive voltage.

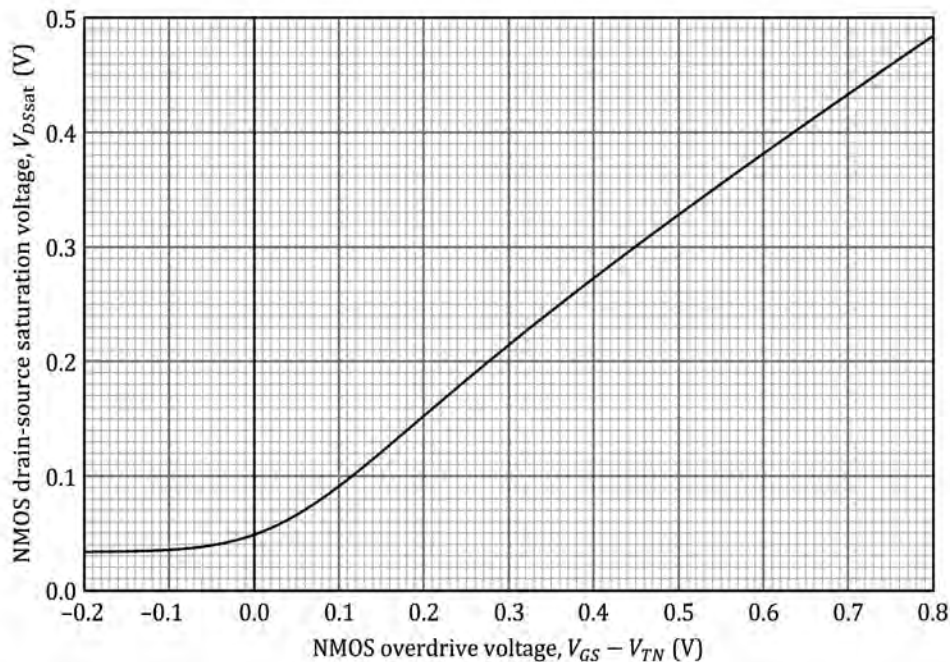


FIGURE 3.8 NMOS drain-source saturation voltage V_{DSsat} as a function of overdrive voltage $V_{GS} - V_{TN}$.

3.3.4 Sheet Current 方块电流

The dc current I_D is yet another metric of device performance. Considering the saturation mode of operation needed in typical applications, the long-channel model prescribes (3.16) or (3.24) for I_D . According to these expressions, drain current is controlled by three design variables, namely, channel width, channel length, and overdrive voltage. But if we normalize the current to aspect ratio W/L , the outcome

$$I_{\square} \triangleq \frac{I_D}{W/L} \quad (3.25)$$

will contain only the overdrive voltage as a design variable. Although $I_{\square}$ also depends weakly on the drain-source voltage, the consequences of this dependence are handled by the output resistance parameter r_{ds} as will be explained in Subsection 3.3.6. In this book, we refer to $I_{\square}$ as *sheet current*, and quantify it in units of A/sq (amperes per square) because it represents the current of one square channel segment of dimensions $L \times L$ as illustrated in Fig. 3.9 on an exemplar MOSFET channel of $W/L = 6.5$. Since there are altogether 6.5 squares along the width, the drain current of this example is given by $I_D = 6.5 \times I_{\square}$.

A single plot of $I_{\square}$ versus $V_{GS} - V_{TN}$ can be used as a common design tool for all devices of a given set of threshold option, process corner, and temperature. This tool is extracted accurately with a single dc-sweep analysis run with the input file `nmos2.sp` described in Subsection 3.3.3. In addition to selecting typical channel dimensions, setting $V_{SB} = 0$, and sweeping V_{GS} in accordance with the comments made in the same section, we need to make sure that the device remains in saturation for the entire range of sweep. Since V_{DSsat} is already available to us as a function of overdrive voltage from Fig. 3.8, we can read its maximum value within the range of sweep, and set V_{DS} at least several hundred millivolts larger than this maximum.

Note that the effective aspect ratio W/L appearing in (3.25) is generally different from the *drawn aspect ratio* W_{DR}/L_{DR} . Although the difference is likely to be small for the

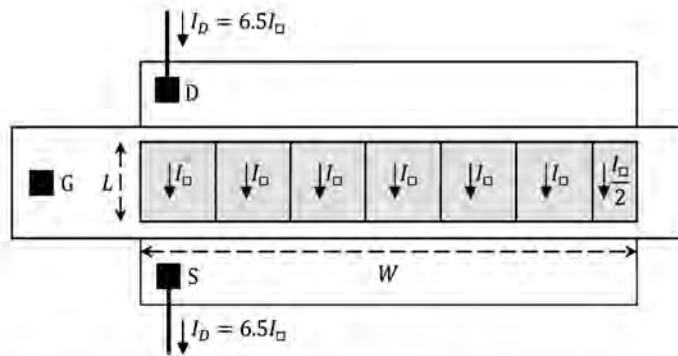


FIGURE 3.9 Sheet current $I_{\square}$ is defined as the fraction of I_D flowing in one square channel segment of dimensions $L \times L$. The channel, shaded in gray, comprises 6.5 squares in this example.

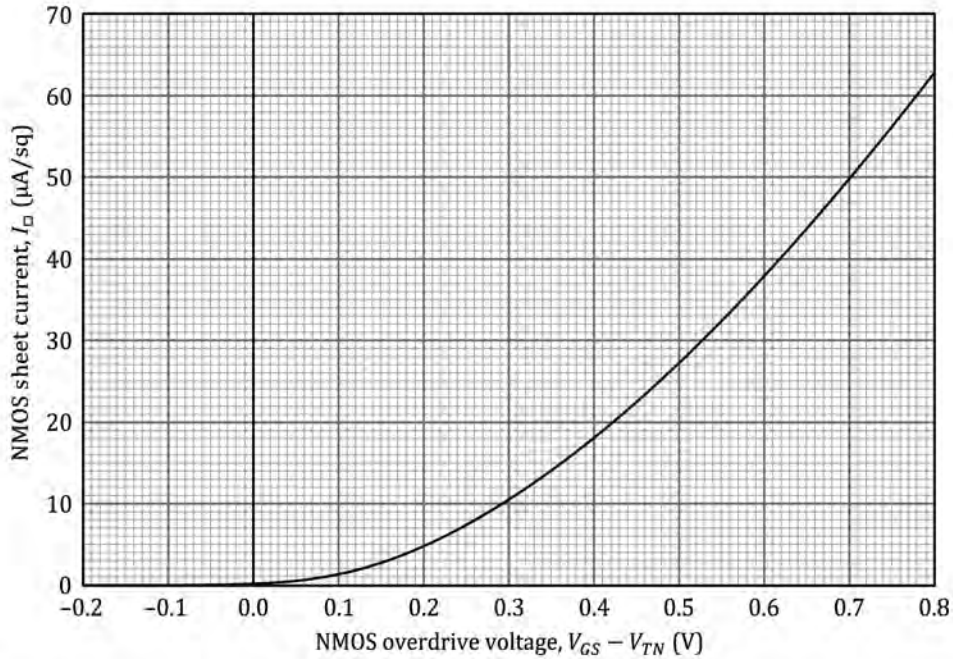


FIGURE 3.10 NMOS sheet current I_D as a function of overdrive voltage $V_{GS} - V_{TN}$.

relatively long and wide device used in simulation, it is prudent to calculate W/L from (3.1) and (3.2), and use the result for normalization. Typing

```
plot @m[id]/7.81 vs v(1,3)-@m[vth]
```

will generate the data shown in Fig. 3.10. The quadratic dependence of I_D on the overdrive voltage, as predicted by the strong-inversion model (3.16), is qualitatively evident from the plot for relatively larger values of the overdrive voltage. However, the prediction that the current should vanish for $V_{GS} - V_{TN} \leq 0$ is not fully corroborated by the simulated plot, though the scale of the vertical axis makes it difficult to detect the residual current. This is where the subthreshold model comes into play. In order to make the subthreshold current visible, the very same simulation result is plotted in Fig. 3.11 using a logarithmic scale for I_D .³ Now, the exponential dependence of drain current on the overdrive voltage, as predicted by the subthreshold model equation (3.24), is obvious for relatively large negative overdrive voltages. Also obvious is the ability of the BSIM model to smoothly bridge the strong-inversion and subthreshold segments of the plot to represent the weak-inversion region in between the two.

It is obvious from (3.25) that a large I_D is desirable for implementing the specified value of I_D with a small aspect ratio W/L , hence a smaller device. As indicated by Fig. 3.10 or Fig. 3.11, this is possible only by adopting a large overdrive voltage. However, it is worth

³For a logarithmic I_D -axis, plot with the command `plot @m[id]/7.81 vs v(1,3)-@m[vth] ylog`.

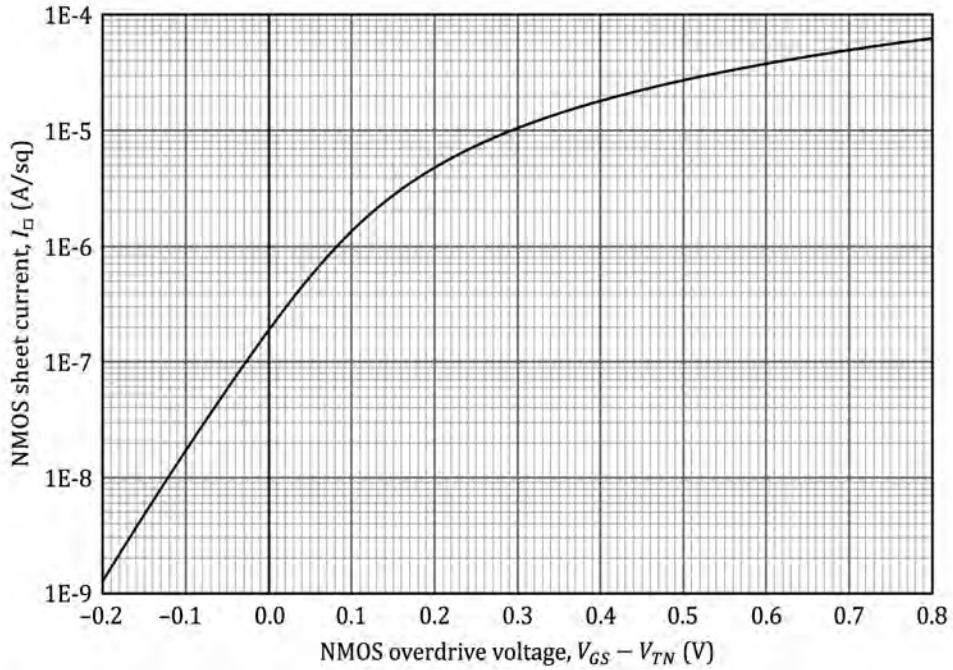


FIGURE 3.11 NMOS sheet current I_D as a function of overdrive voltage $V_{GS} - V_{TN}$. Notice the logarithmic scaling of the vertical axis.

remembering from Subsection 3.3.3 that a large overdrive voltage is counterproductive in terms of V_{DSsat} .

3.3.5 Transconductance Efficiency 跨导效率

Gate Transconductance

As explained in Subsection 3.2.2, g_m is the main performance metric in transconductor applications. Considering the definition (3.5), predictions of the long-channel model for this metric can be formulated by differentiating the saturation-mode drain-current expression (3.16) or (3.24) with respect to v_{GS} . The result is

$$g_m = \frac{W}{L} \frac{\mu_N C_{OX}}{A_N} (V_{GS} - V_{TN}) \quad (\text{in strong inversion}) \quad (3.26)$$

and

$$g_m = \frac{W}{L} \mu_N C_{OX} \frac{A_N - 1}{A_N} \frac{kT}{q} \exp\left(\frac{q}{kT} \frac{V_{GS} - V_{TN}}{A_N}\right) \quad (\text{in subthreshold}). \quad (3.27)$$

According to these two expressions, g_m is not only proportional to aspect ratio W/L but also an increasing function of overdrive voltage in both modes of operation. Furthermore, it depends on numerous physical and technological parameters. Normalizing the transconductance with respect to the drain current I_D flowing under bias conditions greatly reduces this complexity and creates a very useful design tool. Normalization is done by dividing (3.26) by (3.16), and (3.27) by (3.24). The outcome

$$\frac{g_m}{I_D} = \frac{2}{V_{GS} - V_{TN}} \quad (\text{in strong inversion}) \quad (3.28)$$

or

$$\frac{g_m}{I_D} = \frac{1}{A_N} \frac{q}{kT} \quad (\text{in subthreshold}) \quad (3.29)$$

is a measure of how efficiently the drain current is utilized for implementing a specified transconductance, and is appropriately called *transconductance efficiency*. According to (3.28) and (3.29), g_m/I_D (V^{-1}) is inversely proportional to the overdrive voltage in strong inversion and a constant in subthreshold. In fact, it is also a function of channel dimensions, V_{DS} and V_{SB} but for the typical dimensions and overdrive voltage involved in analog applications, these dependencies are secondary. A single plot of g_m/I_D versus $V_{GS} - V_{TN}$ can therefore be used as a common design tool for all devices of a given set of threshold option, process corner, and temperature. This design tool can be extracted accurately with a single dc-sweep analysis run with the input file `nmos2.sp` described in Subsection 3.3.3. Typing

```
plot @m[gm]/@m[id] vs v(1,3)-@m[vth]
```

will generate the plot shown in Fig. 3.12 for our exemplar technology.

The prediction of (3.28) that g_m/I_D is inversely proportional to overdrive in strong inversion is verified by the plot. Also obvious is the saturation tendency of g_m/I_D as predicted by (3.29) for subthreshold operation. Full constancy can be observed only in deep-subthreshold operation beyond the overdrive range displayed in Fig. 3.12.

A large g_m/I_D enables us to realize a given transconductance with less current, hence less power. Obviously, lowering the overdrive is the only way in which this can be accomplished.

Bulk Transconductance and Total Transconductance

Bulk transconductance g_{mb} , as defined by (3.6), is a metric of the sensitivity of drain current to source-bulk voltage. Since the variation of i_D with v_{SB} is due mainly to the dependence of V_{TN} on v_{SB} , (3.6) for NMOS can be rewritten as

$$g_{mb} = - \frac{\partial i_D}{\partial V_{TN}} \frac{dV_{TN}}{dv_{SB}} \bigg|_{v_{SB}=V_{SB}} \quad (3.30)$$

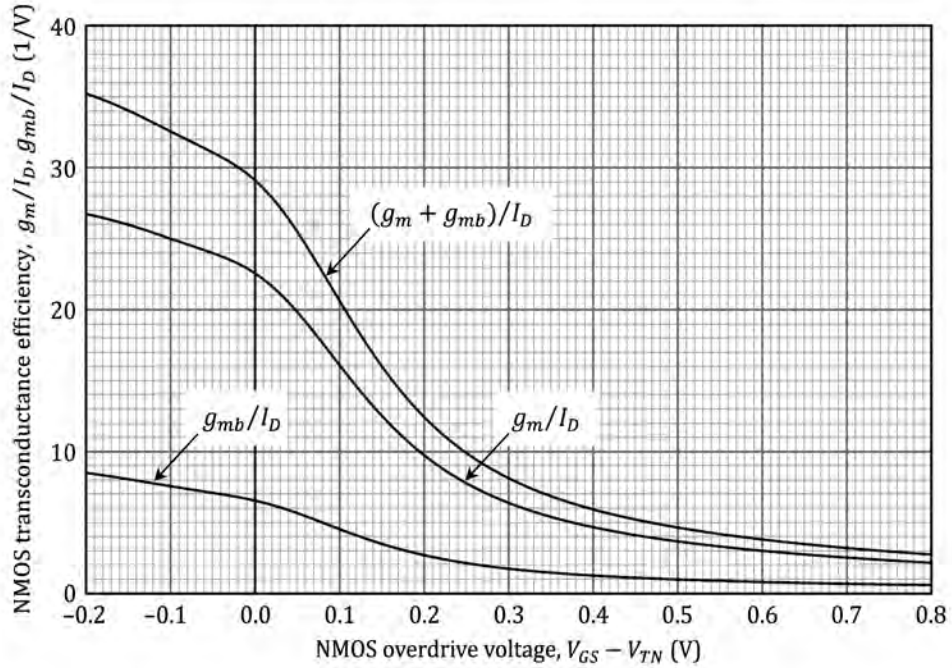


FIGURE 3.12 NMOS transconductance efficiency g_m/I_D and bulk-transconductance efficiency g_{mb}/I_D as functions of overdrive voltage $V_{GS} - V_{TN}$. Also shown is the plot of the sum $(g_m + g_{mb})/I_D$.

Inspecting (3.16) and (3.24), and considering (3.5), we reach the conclusion $\partial i_D / \partial V_{TN} = -\partial i_D / \partial v_{GS} = -g_m$, which turns (3.30) into

$$g_{mb} = g_m \left. \frac{dV_{TN}}{dv_{SB}} \right|_{v_{SB}=V_{SB}}. \quad (3.31)$$

Differentiating (3.10) with respect to v_{SB} , and substituting the outcome into (3.31), and comparing the outcome with (3.14), we finally arrive at

$$g_{mb} = (A_N - 1)g_m. \quad (3.32)$$

Although we used the more accurate expression (3.14) of A_N to arrive at (3.32), we will continue treating A_N in (3.32) as a constant described by (3.15). In our exemplar technology, it equals 1.31, and therefore indicates $g_{mb} = 0.31 \times g_m$ from (3.32). Obviously, g_{mb} is a scaled-down copy of g_m , which is why *bulk-transconductance efficiency* is also a fixed fraction of the transconductance efficiency as described by

$$\frac{g_{mb}}{I_D} = (A_N - 1) \frac{g_m}{I_D}. \quad (3.33)$$

As will be explained in Section 3.7, g_{mb} is involved in device operation only if v_{SB} varies with time. For example, if the source terminal of M6, M7, or M8 of Fig. 3.3(c) carries a variable voltage with respect to the bulk at V_{SS} , then, g_{mb} becomes effective on device performance. It is definitely ineffective in M5 because this device operates with $v_{SB} = 0$. It will also be ineffective in a device with a nonzero but constant V_{SB} . Also note that, in those applications where both v_{GS} and v_{SB} vary, the sum g_{mb} and g_m affects device operation. Called *total-transconductance efficiency*, the normalized sum of the two is given by

$$\frac{g_m + g_{mb}}{I_D} = A_N \frac{g_m}{I_D} \quad (3.34)$$

Spice input file `nmos2.sp` as described in Subsection 3.3.3 is also good for extracting bulk-transconductance efficiency. The command

```
plot @m[gmbs]/@m[id] vs v(1,3)-@m[vth]
```

will yield the plot marked g_{mb}/I_D in Fig. 3.12, where $(g_m + g_{mb})/I_D$ is also plotted.

3.3.6 Output Resistance and Early Voltage 输出阻抗和厄利电压

The long-channel models presented in Subsection 3.3.1 give us no clue about the variables of the output resistance r_{ds} for a device in saturation. Both (3.16) and (3.24) predict a drain current independent of v_{DS} , hence $r_{ds} = \infty$. In practice, r_{ds} is a very complex function of W_{DR} , L_{DR} , $V_{DS} - V_{DSSat}$, and $V_{GS} - V_{TN}$. Since all these variables have comparable influence, we cannot develop a general design tool of a single variable for r_{ds} unlike what we have already done for V_{DSSat} , $I_{\square}$, and transconductance efficiencies. Any device whose specifications include r_{ds} must therefore be designed by iterative simulation without relying on a general tool. One simplifying opportunity, however, is offered by the fact that the so-called *Early voltage* defined by

$$V_A \doteq r_{ds} I_D \quad (3.35)$$

is independent of W_{DR} .⁴ By specifying V_A instead of r_{ds} , the number of variables involved in design is thus reduced to three. We now discuss the way in which V_A varies with each of the three.

$V_{DS} - V_{DSSat}$ is a measure of how deep a device is in saturation. Shown in Fig. 3.13 is the variation of V_A with $V_{DS} - V_{DSSat}$ for eight different values of $V_{GS} - V_{TN}$ ranging from -0.2 V in subthreshold to 0.5 V in strong inversion. These plots are extracted with simulation for $L_{DR} = 0.8 \mu\text{m}$ in our exemplar technology. Before starting to review them, note that there are typically three secondary effects that prevent r_{ds} and therefore V_A from being infinitely high in saturation: (a) Channel-length modulation (CLM), which

⁴The term “Early voltage” is also a BJT inheritance. It is named after James M. Early who was first to describe the physical cause of the finite output resistance of BJT in 1952.

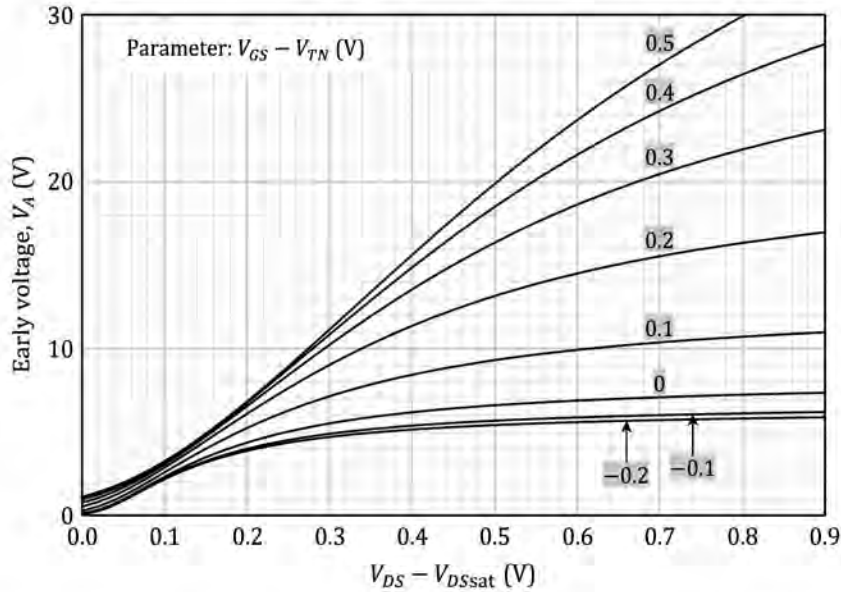


FIGURE 3.13 Early voltage V_A as a function of depth of saturation $V_{DS} - V_{DSsat}$ and overdrive voltage $V_{GS} - V_{TN}$ in an NMOS device of $L_{DR} = 0.8 \mu\text{m}$.

is dominant for small values of $V_{DS} - V_{DSsat}$; (b) drain-induced barrier lowering (DIBL), which is dominant for small values of $V_{GS} - V_{TN}$; and (c) substrate-current-induced body bias (SCBE), which is dominant for large values of $V_{DS} - V_{DSsat}$. Now, let us return to Fig. 3.13, where we make the following three observations:

- V_A is small and relatively independent of the overdrive voltage for small values of $V_{DS} - V_{DSsat}$ lesser than about 150 mV. This is where CLM is dominant.
- For higher values of $V_{DS} - V_{DSsat}$, the device gradually escapes from CLM, and therefore V_A improves. However, the extent of improvement depends strongly on the overdrive voltage due to the dominance of DIBL. Obviously the higher the overdrive, the larger the Early voltage.
- SCBE is nowhere near effective within the range of $V_{DS} - V_{DSsat}$ in Fig. 3.13. Had it been effective, the plots would decline at the high end of $V_{DS} - V_{DSsat}$.

According to these observations, raising $V_{DS} - V_{DSsat}$ to escape from CLM and adopting a large $V_{GS} - V_{TN}$ to suppress DIBL are beneficial for a large V_A . Since V_{DS} is not an influential variable of any of the performance metrics discussed previously, we indeed manipulate it for raising V_A as long as it is not restricted by topological considerations. Treating $V_{GS} - V_{TN}$ as a design variable for V_A , however, is hardly practical because overdrive is the sole variable appearing in all design tools we have developed so far for other device metrics except for the threshold voltage. Only in rare cases, we can manipulate it for achieving a large V_A without adversely affecting one or more of these metrics. Even in such cases, we need to be aware of the fact that a large overdrive voltage leads to a

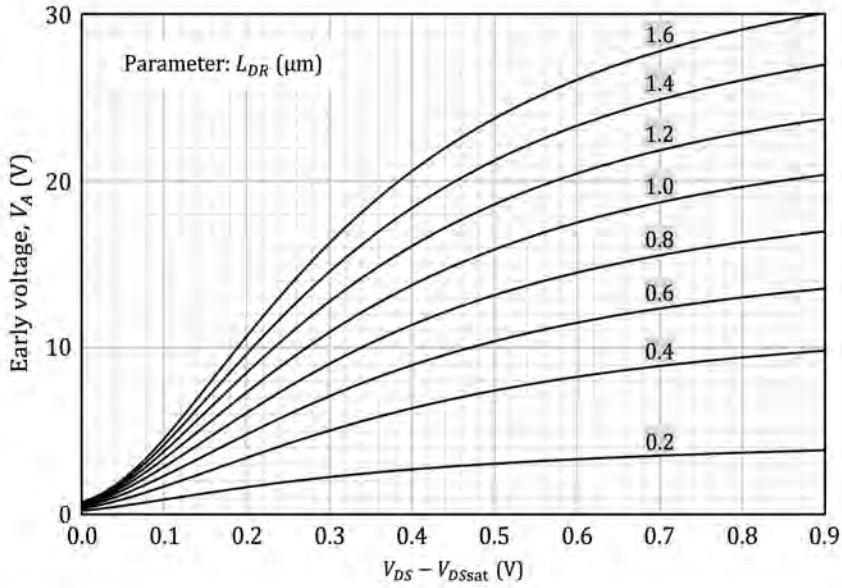


FIGURE 3.14 Early voltage V_A as a function of $V_{DS} - V_{DSsat}$ and L_{DR} in an NMOS device of $V_{GS} - V_{TN} = 0.2$ V.

large V_{DSsat} , hence a small $V_{DS} - V_{DSsat}$ for a given V_{DS} . Unless we compensate for this loss by raising V_{DS} , V_A may actually decrease rather than increase.

Finally, we move to the effect of channel length on V_A . Shown in Fig. 3.14 is the simulated variation of V_A with $V_{DS} - V_{DSsat}$ in our exemplar technology for eight different values of L_{DR} starting with the minimum permissible $L_{DR(min)} = 0.2$ μm . Overdrive voltage is fixed at 0.2 V. Obviously, V_A is a strongly increasing function of the channel length, which not only explains why minimum-geometry MOSFETs are rarely used in analog circuits but also indicates L_{DR} as the prime design variable of V_A .

Example 3.2 An NMOS transistor is specified by $W_{DR} = 20$ μm and $L_{DR} = 0.9$ μm . Determine its drain current for the bias condition $V_{GS} = 0.55$ V, $V_{DS} = 0.9$ V, and $V_{SB} = 0$ by using (a) long-channel model equations, (b) design tools if applicable, and (c) Spice operation-point analysis.

Work First, calculate the technological constants appearing in long-channel model equations:

- (3.18) $\Rightarrow C_{OX} = \frac{3.5 \times 10^{-11}}{4.1 \times 10^{-9}} = 8.54 \times 10^{-3}$ F/m².
- (3.11) $\Rightarrow \phi_{FB} = 0.0258 \times \ln \frac{2.35 \times 10^{17}}{1.1 \times 10^{10}} = 0.435$ V.
- (3.15) $\Rightarrow A_N = 1 + \frac{0.578}{2\sqrt{2 \times 0.435}} = 1.31$.

Next, calculate effective dimensions:

- (3.1) $\Rightarrow W = 20 \times 10^{-6} - 1 \times 10^{-8} - 2 \times 2.88 \times 10^{-9} = 2 \times 10^{-5} \text{ m.}$
- (3.2) $\Rightarrow L = 0.9 \times 10^{-6} + 0 - 2 \times 1.71 \times 10^{-8} = 8.66 \times 10^{-7} \text{ m.}$

Now determine I_D :

- From long-channel model equations:
 - (3.10) $\Rightarrow V_{TN} = V_{TH0} = 0.37 \text{ V}$ because $V_{SB} = 0$. $V_{GS} - V_{TN} = 0.18 \text{ V}$. According to (3.8) device is not in cutoff.
 - (3.13) $\Rightarrow V_{DSSat} = \frac{0.55 - 0.37}{1.31} = 0.137 \text{ V}$, which is less than $V_{DS} = 0.9 \text{ V}$. Therefore (3.12) implies saturation.
 - (3.17) $\Rightarrow \mu_N = \frac{293 \times 10^{-4}}{1 - 1.22 \times 10^{-9} \frac{2 \times 0.37}{4.1 \times 10^{-9}} + 2.32 \times 10^{-18} \left(\frac{2 \times 0.37}{4.1 \times 10^{-9}} \right)^2} = 3.42 \times 10^{-2} \text{ m}^2/\text{Vs.}$
 - (3.16) $\Rightarrow I_D = \frac{2 \times 10^{-5}}{8.66 \times 10^{-7}} \frac{3.42 \times 10^{-2} \times 8.54 \times 10^{-3}}{2 \times 1.31} 0.18^2 = 8.34 \times 10^{-5} \text{ A (83.4 } \mu\text{A).}$
- From design tools:
 - Figure 3.7 $\Rightarrow V_{TN} = 0.425 \text{ V} \Rightarrow V_{GS} - V_{TN} = 0.125 \text{ V}$.
 - Figure 3.8 $\Rightarrow V_{DSSat} = 0.105 \text{ V}$, which is less than $V_{DS} = 0.9 \text{ V}$. Therefore (3.12) implies saturation.
 - Figures 3.10 and 3.11 are applicable because device is in saturation. They indicate $I_{\square} = 2 \times 10^{-6} \text{ A/sq.}$
 - (3.25) $\Rightarrow I_D = 2 \times 10^{-6} \frac{2 \times 10^{-5}}{8.66 \times 10^{-7}} = 4.62 \times 10^{-5} \text{ A (46.2 } \mu\text{A).}$
- Spice operation-point analysis: Use the following input file.

```
filename: ex3-2.sp
.incl bsim.sp
m 2 1 0 0 cmosn w=20u l=0.9u
vds 2 0 0.9
vgs 1 0 0.55
.op
.end
```

The command show yields a list of all variables and parameters including the drain current $I_D = 44.8 \mu\text{A}$.

The reader is invited to take this opportunity to compare graphical design tools with long-channel model in (a) the volume of calculation involved, and (b) the accuracy of prediction offered. ▲

3.4 PMOS Design Relations and Tools PMOS晶体管设计方程和工具

The PMOS counterparts of long-channel model equations, design relations, and tools are presented in this section. Sign convention will be consistent with the references set up in Fig. 3.2 for PMOS devices.

3.4.1 Strong-Inversion Model 强反型模型

Cutoff prevails for

$$v_{GS} - V_{TP} \geq 0 \text{ (condition of cutoff),} \quad (3.36)$$

and implies

$$i_D = 0 \text{ (in cutoff).} \quad (3.37)$$

PMOS threshold voltage is given by

$$V_{TP} = V_{TH0} - K1(\sqrt{2\phi_{FB} - v_{SB}} - \sqrt{2\phi_{FB}}), \quad (3.38)$$

where

$$\phi_{FB} = \frac{kT}{q} \ln \frac{N_{CH}}{n_i}. \quad (3.39)$$

According to (3.38), V_{TP} equals V_{TH0} for $v_{SB} = 0$. As listed in Table 3.1, V_{TH0} equals -0.38 V in our exemplar technology. Also note from (3.38) that V_{TP} varies in the negative direction due to body effect as v_{SB} is made increasingly negative.

Saturation prevails for

$$v_{GS} - V_{TP} \leq 0 \text{ and } v_{DS} \leq V_{DSsat} \text{ (condition of saturation),} \quad (3.40)$$

where V_{DSsat} is given by

$$V_{DSsat} = \frac{v_{GS} - V_{TP}}{A_p}, \quad (3.41)$$

and A_p is defined by

$$A_p \triangleq 1 + \frac{K1}{2\sqrt{2\phi_{FB} - v_{SB}}}. \quad (3.42)$$

To keep the model simple, we usually ignore the dependence of A_p on v_{SB} , and simplify (3.42) into

$$A_p = 1 + \frac{K1}{2\sqrt{2\phi_{FB}}}. \quad (3.43)$$

In saturation, drain current is described by

$$i_D = \frac{W}{L} \frac{\mu_p C_{OX}}{2A_p} (v_{GS} - V_{TP})^2 \quad (\text{in saturation}), \quad (3.44)$$

where hole mobility μ_p (m^2/Vs) is given approximately by

$$\mu_p \approx \frac{U_0}{1 + U_A \frac{|2 \times V_{TH0}|}{TOX} + U_B \left(\frac{|2 \times V_{TH0}|}{TOX} \right)^2}, \quad (3.45)$$

and C_{OX} is described by

$$C_{OX} \doteq \frac{\epsilon_{ox}}{TOX}. \quad (3.46)$$

Nonsaturation prevails for

$$v_{GS} - V_{TP} \leq 0 \text{ and } v_{DS} \geq V_{DSsat} \quad (\text{condition of nonsaturation}), \quad (3.47)$$

Drain current in nonsaturation is described by

$$i_D = \frac{W}{L} \mu_p C_{OX} \left(v_{GS} - V_{TP} - \frac{A_p}{2} v_{DS} \right) v_{DS} \quad (\text{in nonsaturation}). \quad (3.48)$$

Drain current, as predicted by the strong-inversion model, is plotted in Fig. 3.15(a) as a function of overdrive voltage $v_{GS} - V_{TP}$ for a fixed drain-source voltage V_{DS} . Its variation with v_{DS} for a fixed negative overdrive voltage $V_{GS} - V_{TP}$ is shown in Fig. 3.15(b).

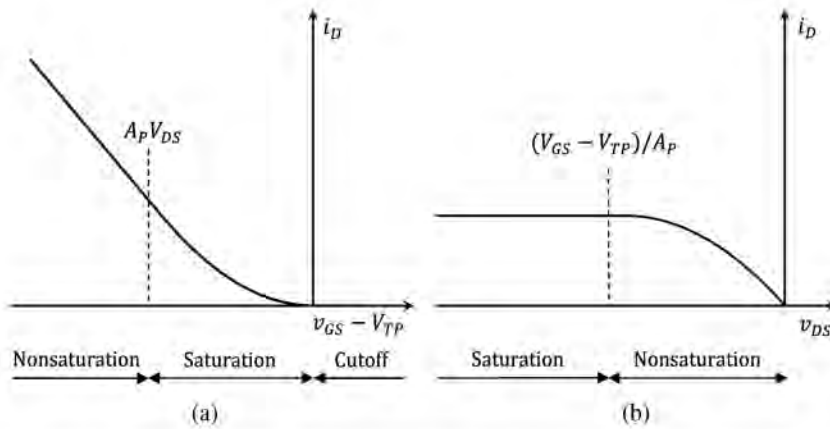


FIGURE 3.15 Strong-inversion model plots of PMOS drain-current versus (a) overdrive voltage and (b) drain-source voltage.

3.4.2 Subthreshold Model 亚阈值模型

The subthreshold drain current of a PMOS device is modeled by

$$i_D = \frac{W}{L} \mu_P C_{OX} (A_P - 1) \left(\frac{kT}{q} \right)^2 \exp \left(-\frac{q}{kT} \frac{v_{GS} - V_{TP}}{A_P} \right) \left[1 - \exp \left(\frac{q}{kT} v_{DS} \right) \right]. \quad (3.49)$$

Saturation in subthreshold is defined by

$$v_{GS} - V_{TP} \geq 0 \text{ and } v_{DS} \leq V_{DSsat} \text{ (condition of saturation),} \quad (3.50)$$

where V_{DSsat} is given by

$$V_{DSsat} \approx -2 \frac{kT}{q}. \quad (3.51)$$

In saturation, (3.49) is simplified into

$$i_D = \frac{W}{L} \mu_P C_{OX} (A_P - 1) \left(\frac{kT}{q} \right)^2 \exp \left(-\frac{q}{kT} \frac{v_{GS} - V_{TP}}{A_P} \right) \text{ (in saturation),} \quad (3.52)$$

3.4.3 Threshold Voltage 阈值电压

The following Spice input file can be used for extracting and plotting the design tool relating V_{TP} to V_{SB} and L_{DR} .

```
filename: pmos1.sp (pmos vth)
.incl bsim.sp
m 2 1 3 0 cmosp w=6u l=0.2u
vds 2 3 -0.9
vgs 1 3 -0.9
vsb 3 0 0
.save @m[vth]
.control
destroy all
foreach length 0.2u 0.4u 0.6u 0.8u 1.2u 2u 4u 8u
alter @m[1]=$length
dc vsb -1.2 0 20m
end
plot (-dc1.@m[vth]) (-dc2.@m[vth])
+ (-dc3.@m[vth]) (-dc4.@m[vth]) (-dc5.@m[vth])
+ (-dc6.@m[vth]) (-dc7.@m[vth]) (-dc8.@m[vth])
.endc
.end
```

The resulting plots are shown in Fig. 3.16.

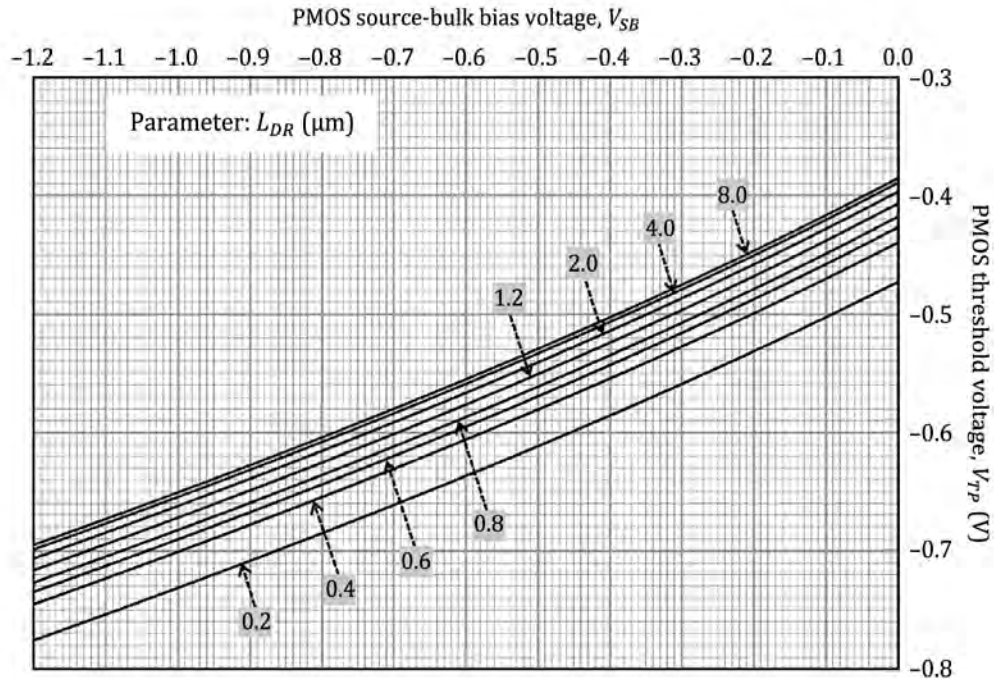


FIGURE 3.16 PMOS threshold voltage V_{TP} as a function of source-bulk bias voltage V_{SB} and drawn channel length L_{DR} .

3.4.4 Drain-Source Saturation Voltage, Sheet Current, and Transconductance Efficiency 漏源饱和电压、方块电流和跨导效率

The following Spice input file is appropriate for extracting the design tools relating V_{DSsat} , $I_{\square}$, g_m/I_D , and g_{mb}/I_D to overdrive voltage $V_{GS} - V_{TP}$ for a PMOS device.

```
filename: pmos2.sp (pmos vdsat, id, gm, gmbs)
.incl bsim.sp
m 2 1 3 0 cmosp w=6u l=0.8u
vds 2 3 -0.9
vgs 1 3
vsb 3 0 0
.dc vgs -1.3 -0.2 10m
.save @m[vth] @m[vdsat] @m[id] @m[gm] @m[gmbs] v(1,3)
.end
```

The resulting plots are given in Figs. 3.17 to 3.20. For plotting the variables in accordance with our sign convention, we must take into account the fact that V_{DSsat} and V_{TP} are saved in absolute values in Spice. Proper commands are as follows:

```
 $V_{DSsat}$ : plot -@m[vdsat] vs v(1,3)+@m[vth]
 $I_{\square}$ : plot @m[id]/8.08 vs v(1,3)+@m[vth]
 $g_m/I_D$ : plot @m[gm]/@m[id] vs v(1,3)+@m[vth]
 $g_{mb}/I_D$ : plot @m[gmbs]/@m[id] vs v(1,3)+@m[vth]
```

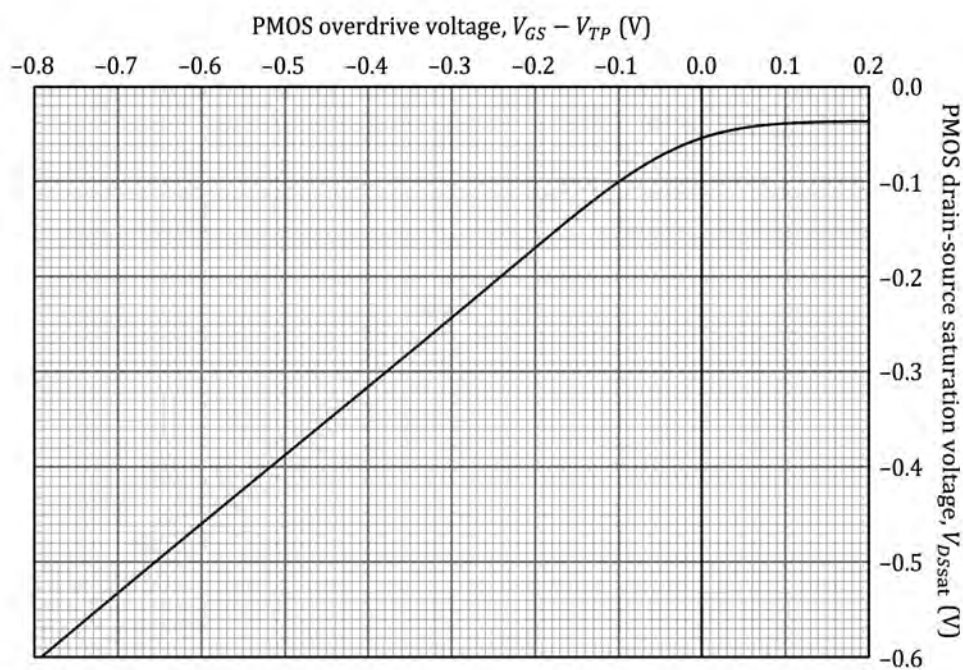


FIGURE 3.17 PMOS drain-source saturation voltage V_{DSsat} as a function of overdrive voltage $V_{GS} - V_{TP}$.

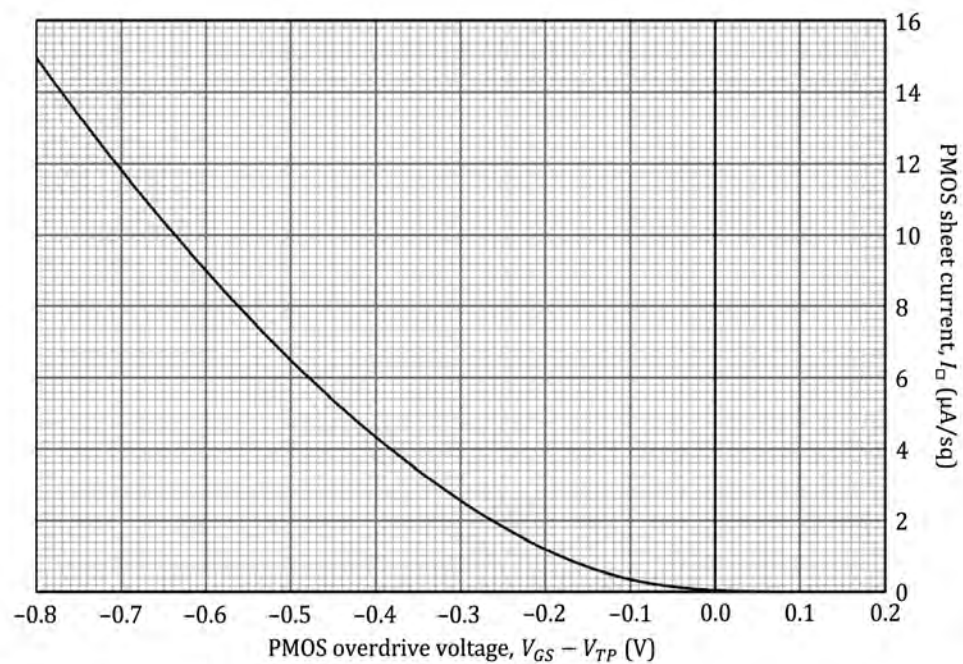


FIGURE 3.18 PMOS sheet current $I_{\square}$ as a function of overdrive voltage $V_{GS} - V_{TP}$.

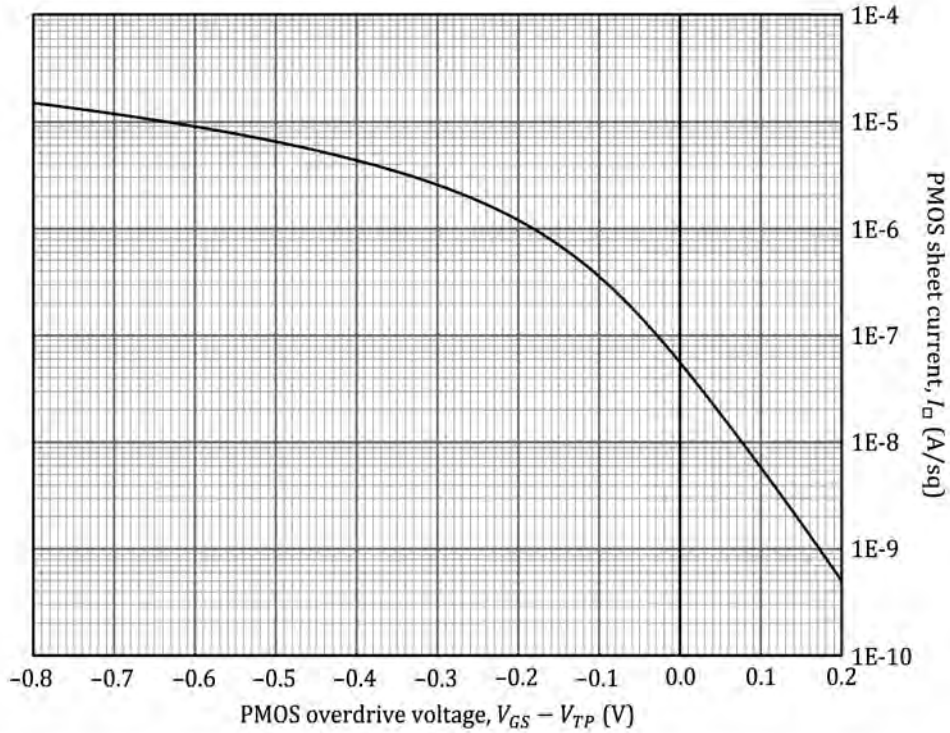


FIGURE 3.19 PMOS sheet current I_{sq} as a function of overdrive voltage $V_{GS} - V_{TP}$. Notice the logarithmic calibration of the vertical axis.

3.4.5 Output Resistance and Early Voltage 输出阻抗和厄利电压

All conclusions reached in Subsection 3.3.6 for the output resistance r_{ds} and the Early voltage V_A are also valid for PMOS devices. In summary, V_A is more convenient to handle than r_{ds} due to its independence of W but an iterative simulation is still needed for designing a PMOS device specified for a particular value of V_A .

Example 3.3 A PMOS transistor is specified by $W_{DR} = 10 \mu\text{m}$ and $L_{DR} = 1 \mu\text{m}$. Its dc terminal voltages with respect to circuit ground are as follows: Bulk $V_B = 1.8 \text{ V}$, source $V_S = 1.4 \text{ V}$, drain $V_D = 0.9 \text{ V}$, gate $V_G = 0.6 \text{ V}$. Determine its drain current by using (a) long-channel model equations, (b) design tools if applicable, and (c) Spice operation-point simulation.

Work First, calculate the technological constants appearing in long-channel model equations:

- (3.46) $\Rightarrow C_{OX} = \frac{3.5 \times 10^{-11}}{4.1 \times 10^{-9}} = 8.54 \times 10^{-3} \text{ F/m}^2$.
- (3.39) $\Rightarrow \phi_{FB} = 0.0258 \times \ln \frac{4.16 \times 10^{17}}{1.1 \times 10^{10}} = 0.450 \text{ V}$.
- (3.43) $\Rightarrow A_p = 1 + \frac{0.572}{2\sqrt{2 \times 0.450}} = 1.30$.

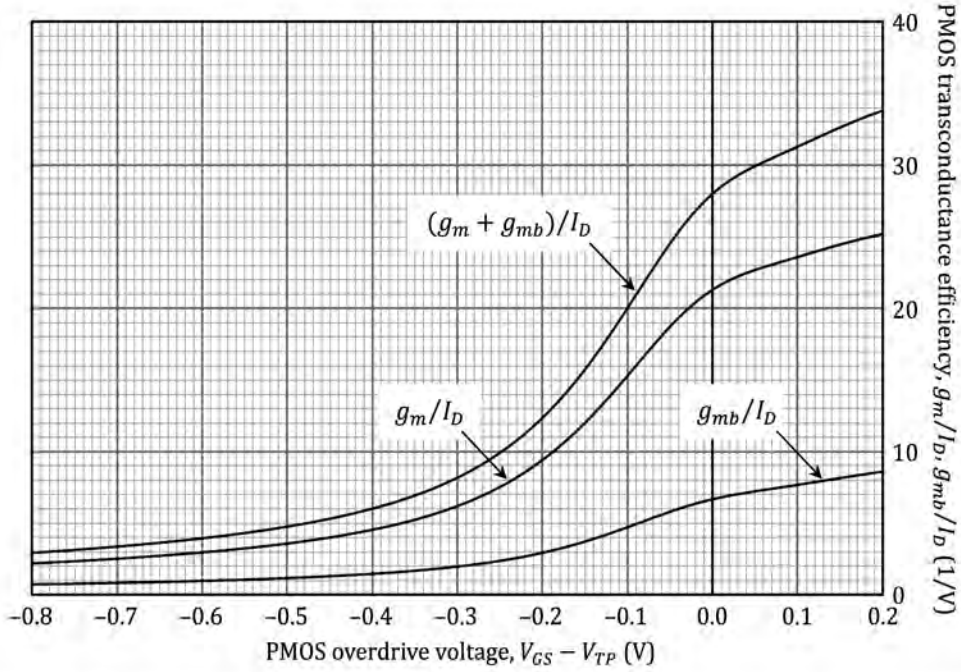


FIGURE 3.20 PMOS transconductance efficiency g_m/I_D and bulk-transconductance efficiency g_{mb}/I_D as functions of overdrive voltage $V_{GS} - V_{TP}$. Also shown is the plot of the sum $(g_m + g_{mb})/I_D$.

Next, calculate effective dimensions and port voltages:

- (3.1) $\Rightarrow W = 10 \times 10^{-6} - 1 \times 10^{-8} - 2 \times 0 = 9.99 \times 10^{-6}$ m.
- (3.2) $\Rightarrow L = 1 \times 10^{-6} - 0 - 2 \times 2.94 \times 10^{-8} = 9.41 \times 10^{-7}$ m.
- $V_{GS} = 0.6 - 1.4 = -0.8$ V, $V_{DS} = 0.9 - 1.4 = -0.5$ V, $V_{SB} = 1.4 - 1.8 = -0.4$ V.

Now determine I_D :

- From long-channel model equations:
 - (3.38) $\Rightarrow V_{TP} = -0.38 - 0.57 \left(\sqrt{2 \times 0.45 - (-0.4)} - \sqrt{2 \times 0.45} \right) = -0.489$ V.
 - $V_{GS} - V_{TP} = -0.8 - (-0.489) = -0.311$ V, which rules out cutoff.
 - (3.41) $\Rightarrow V_{DSsat} = \frac{-0.311}{1.30} = -0.239$ V, which is less negative than $V_{DS} = -0.5$ V. Therefore (3.40) implies saturation.
 - (3.45) $\Rightarrow \mu_p = \frac{109 \times 10^{-4}}{1 + 1.31 \times 10^{-9} \frac{|2 \times (-0.38)|}{4.1 \times 10^{-9}} + 1 \times 10^{-21} \left(\frac{|2 \times (-0.38)|}{4.1 \times 10^{-9}} \right)^2} = 8.77 \times 10^{-3}$ m²/Vs.
 - (3.44) $\Rightarrow I_D = \frac{9.99 \times 10^{-5}}{9.41 \times 10^{-7}} \frac{8.77 \times 10^{-3} \times 8.54 \times 10^{-3}}{2 \times 1.30} (-0.311)^2 = 2.96 \times 10^{-5}$ A (29.6 μ A).

- From design tools:
 - Figure 3.16 $\Rightarrow V_{TP} = -0.535$ V for $V_{SB} = -0.4$ V. Therefore $V_{GS} - V_{TP} = -0.8 - (-0.535) = -0.265$ V.
 - Figure 3.17 $\Rightarrow V_{DSSat} = -0.22$ V, which is less negative than $V_{DS} = -0.5$ V. Therefore (3.40) implies saturation.
 - Figures 3.18 and 3.19 are applicable because device is in saturation. They indicate $I_{\square} = 2$ $\mu\text{A}/\text{sq}$.
 - (3.25) $\Rightarrow I_D = \frac{9.99 \times 10^{-6}}{9.41 \times 10^{-7}} 2 \times 10^{-6} = 21.2$ μA .
- Spice operation-point analysis: Use the following input file.

```
filename: ex3-3.sp
.incl bsim.sp
m 2 1 3 4 cmosp w=10u l=1u
vb 4 0 1.8
vs 3 0 1.4
vd 2 0 0.9
vg 1 0 0.6
.op
.end
```

The command show will display $I_D = 21.6$ μA . ▲

Example 3.4 A MOSFET whose gate is interconnected with one of the current-conducting terminals as shown in Fig. 3.21 finds extensive use in analog circuits. It is called *diode-connected MOSFET*. Using the strong-inversion model, show that the device operates unconditionally in saturation when a positive current i_D is imposed on it.

Work First consider the diode-connected NMOS device. A positive i_D implies a positive v , and thus makes the lower terminal source in accordance with (3.3). The gate terminal being short-circuited to the drain, the device therefore operates with $v_{GS} = v_{DS} = v$. Since a current is imposed, the device cannot be in cutoff. This implies a positive overdrive $v - V_{TN} > 0$, which satisfies the first part of the saturation condition (3.12). As to the second part of (3.12), compare $v_{DS} = v$ with $V_{DSSat} = (v - V_{TN})/A_N$. Since V_{TN} is positive, and A_N is positive and greater than 1, V_{DSSat} is definitely smaller than v_{DS} , hence unconditional saturation.

Now consider the diode-connected PMOS device. A positive i_D implies a positive v , and thus makes the upper terminal source in accordance with (3.3). Gate being shorted to

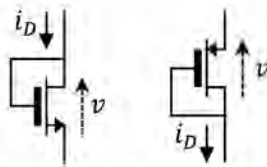


FIGURE 3.21 Diode-connected NMOS and PMOS devices.

the drain, the device therefore operates with $v_{GS} = v_{DS} = -v$. Since current is nonzero, the device cannot be in cutoff. This implies a negative overdrive $-v - V_{TP} < 0$, which satisfies the first part of the saturation condition (3.40). As to the second part of (3.40), compare $v_{DS} = -v$ with $V_{DSsat} = (-v - V_{TP})/A_p$. Since V_{TP} is negative, and A_p is positive and greater than 1, V_{DSsat} is definitely less negative than v_{DS} , hence unconditional saturation.

Actually, a diode-connected MOSFET operates in saturation even under subthreshold conditions as long as v is more positive than a few kT/q in NMOS, or more negative than a few $-kT/q$ in PMOS. See (3.23) and (3.51) for a proof. ▲

Example 3.5 Shown in Fig. 3.22(a) is the schematic of a *common-source* voltage-amplifier stage. M1 operates as a transconductor with its source and bulk terminals tied to circuit ground. M2 operates as a current source with its source and bulk terminals tied to a positive power supply $V_{DD} = 1.8$ V. Dimensions are $W_{DR1} = 6$ μm and $L_{DR1} = 0.8$ μm in M1, and $W_{DR2} = 24$ μm and $L_{DR2} = 0.8$ μm in M2. The gate of M2 is tied to a bias voltage $V_B = 1.25$ V. The gate of M1 receives the input voltage v_i of the amplifier. v_o represents the output voltage. All these voltages are defined with respect to ground. Based on the strong-inversion model, interpret the simulated plot of the *transfer characteristic* (v_o versus v_i) shown in Fig. 3.22(b).

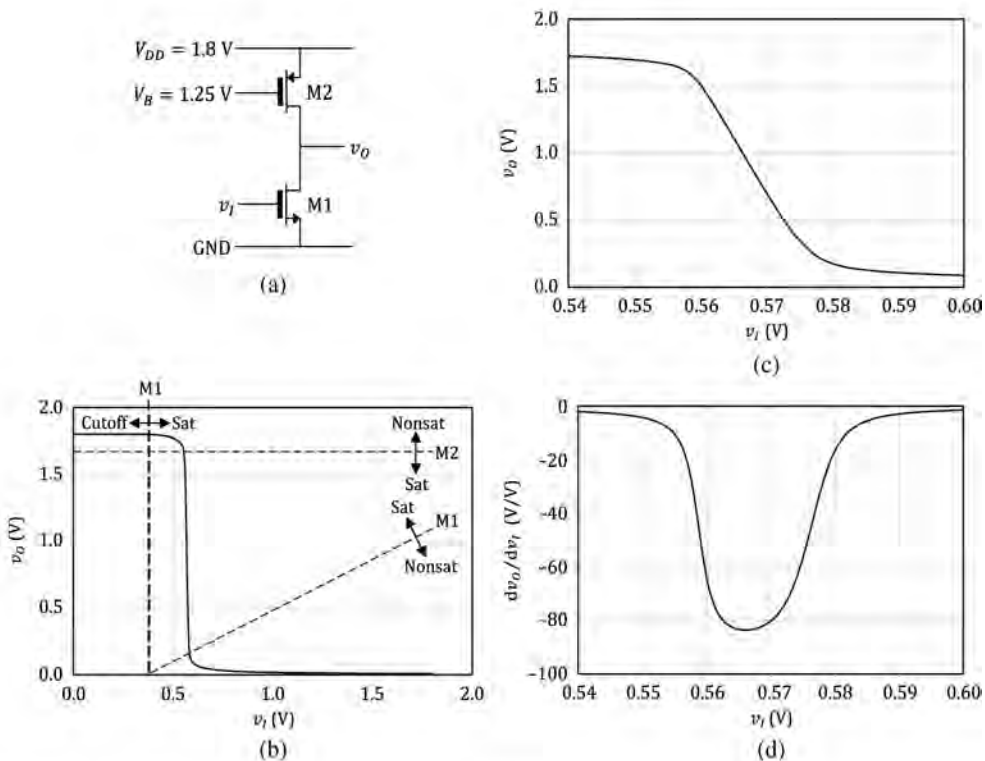


FIGURE 3.22 Common-source amplifier stage of Example 3.5. (a) Electrical schematic. (b) Simulated transfer characteristics (v_o versus v_i). (c) Simulated transfer characteristic zoomed in around the high-gain segment. (d) Simulated slope of the transfer characteristic (dv_o/dv_i versus v_i).

Work In order to interpret the transfer characteristics properly, we need to determine the operation modes of the two devices.

Consider M2 first. This device operates with $v_{SB} = 0$ implying $V_{TP} = V_{TH0} = -0.38$ V from Table 3.1. Its overdrive voltage is fixed at $V_B - V_{DD} - V_{TP} = 1.25 - 1.8 - (-0.38) = -0.17$ V, which, being negative, implies inversion. Whether this device operates in nonsaturation or saturation should be determined by comparing its drain-source voltage $v_{DS2} = v_O - V_{DD} = v_O - 1.8$ with V_{DSsat2} . For the overdrive calculated above, (3.41) yields $V_{DSsat2} = -0.131$ V ($A_p = 1.30$ is borrowed from Example 3.3). Carrying these into (3.40) and (3.47), we conclude that nonsaturation prevails for $v_O > 1.669$ V, and saturation prevails otherwise. The boundary between the two modes of operation is shown with a dashed line in Fig. 3.22(b).

Next, consider the operation modes of M1. This device also operates with $v_{SB} = 0$; hence, a zero-bias threshold voltage $V_{TN} = V_{TH0} = 0.369$ V from Table 3.1. According to (3.8), $v_I < 0.369$ V implies cutoff. Otherwise, the device is in inversion, where saturation and nonsaturation modes are demarcated by the line $v_O = V_{DSsat1} = (v_I - V_{TN})/A_N = (v_I - 0.369)/1.31$. This line is also plotted in Fig. 3.22(b). As indicated by (3.12) and (3.19), saturation prevails above the line, and nonsaturation prevails below the line.

Now we can interpret the characteristic. First, consider the part of the characteristic for $v_I \leq V_{TN} = 0.369$ V where M1 is in cutoff. Since M1 does not conduct a current, the current source M2 cannot conduct a current either. However, since M2 operates with a negative overdrive -0.170 V, its current must somehow vanish outside cutoff! Since a finite overdrive in a saturated device would definitely give rise to a finite current as implied by (3.44), this device shouldn't be in saturation. This leaves nonsaturation as the only possible mode of operation. Inspecting (3.48) of nonsaturation, we indeed observe that a vanishing drain-source voltage rules out current flow. This conclusion implies $v_O = V_{DD} = 1.8$ V, which is indeed what the simulated transfer characteristic shown in Fig. 3.22(b) exhibits for $v_I < 0.369$ V.

Moving v_I above 0.369 V forces M1 into saturation. The current of this device now increases with v_I in accordance with (3.16). Since M2 must conduct the same amount of current while remaining in nonsaturation with a fixed overdrive, its drain-source voltage must grow in the negative direction. As a result, v_O starts declining as observed in Fig. 3.22(b).

As the declining v_O drops below 1.669 V, M2 also enters into saturation. This occurs around $v_I = 0.57$ V, and causes the transfer characteristic to drop off abruptly before M1 goes into nonsaturation. Shortly, we will examine this very important segment in detail. Presently, let us continue with the rest of the transfer characteristics, where M1 moves into nonsaturation with an increasing overdrive voltage $v_I - V_{TN}$ while M2 remains in saturation with a fixed overdrive voltage $V_B - V_{DD} - V_{TP}$ hence a fixed current. The only way in which the nonsaturated M1 can conduct the fixed current of M2 with an increasing overdrive is to reduce its own drain-source voltage v_O . This is why v_O continues to decline as v_I increases beyond 0.57 V.

The near-vertical segment is where this circuit is operated as an amplifier stage. Its slope corresponds to the voltage gain available from the stage. Now, let us attempt to characterize this segment analytically by equating the long-channel model equations of saturation (3.16) and (3.44). The result predicts a perfectly vertical segment located at

$$V_I = \sqrt{\frac{\mu_P A_N W_2 L_1}{\mu_N A_P L_2 W_1} (V_B - V_{DD} - V_{TP})^2} + V_{TN},$$

hence an infinitely large voltage gain at this particular input voltage! In actual fact, the slope of the segment is finite as observed in the close-up shown in Fig. 3.22(c). The derivative of the close-up shown in Fig. 3.22(d) clearly shows that the gain is no larger than -83.6 V/V. The reason for this discrepancy between the predictions of long-channel model and simulation is that the variation of i_D with v_{DS} in saturation is ignored in (3.16) and (3.44). As explained in Subsection 3.3.6, this variation is characterized with the output-resistance parameter r_{ds} . Indeed, the significance of r_{ds} in analog design cannot be overemphasized! ▲

3.5 Thermal Effects 热效应

MOSFET characteristics are generally temperature dependent. In strong inversion, where drift is the underlying carrier motion, this dependence arises mainly from the temperature sensitivity of the threshold voltage and mobility. In subthreshold operation, in which current flows predominantly by diffusion, additional thermal sensitivities arise. In this section, we discuss the effect of temperature on the design tools we have developed for threshold voltage, drain-source saturation voltage, sheet current, and transconductance efficiency. The plots presented originally in Figs. 3.7 through 3.12 for an NMOS device are extracted for room temperature $T = 27^\circ\text{C}$. They are redrawn in Figs. 3.23(a) to (e) for two additional temperatures $T = -40^\circ\text{C}$ and $T = 125^\circ\text{C}$.⁵ The reader would benefit from extracting the PMOS counterparts of these plots with input file `pmos2.sp`.

As observed in Fig. 3.23(a), V_{TN} is a decreasing function of temperature. Plots of V_{DSsat} shown in Fig. 3.23(b), indicate a negligible temperature dependence except for large values of overdrive. In contrast, the variation of sheet current with temperature is quite significant as indicated by Fig. 3.23(c) on linear axis and by Fig. 3.23(d) on logarithmic axis. What is interesting about these plots is that they indicate a positive temperature sensitivity in deep subthreshold but a negative one in strong inversion. The latter is mainly due to the negative temperature sensitivity of the mobility μ_N appearing in (3.16). Although mobility also appears in the subthreshold drain-current expression (3.24), the quadratic dependence on the thermal voltage kT/q overrides its effect in subthreshold, and gives rise to the positive sensitivity observed in these plots. The plots shown in Fig. 3.23(e) indicate no significant temperature sensitivity for transconductance efficiency and bulk-transconductance efficiency in strong inversion. In subthreshold, however, these two become strongly decreasing functions of temperature as predicted by (3.29).

As an interested reader can verify with simulation, the effect of rising temperature on PMOS performance metrics is as follows: V_{TP} becomes less negative, V_{DSsat} becomes more negative in deep subthreshold but less negative in strong inversion, $I_{\square}$ becomes higher in deep subthreshold but lower in strong inversion, g_m/I_D and g_{mb}/I_D become smaller in deep subthreshold and slightly higher in strong inversion.

⁵ For a Spice simulation at any temperature other than $T = 27^\circ\text{C}$, just add a line `.options temp=X` to the input file, where X is the desired simulation temperature in degree centigrade.

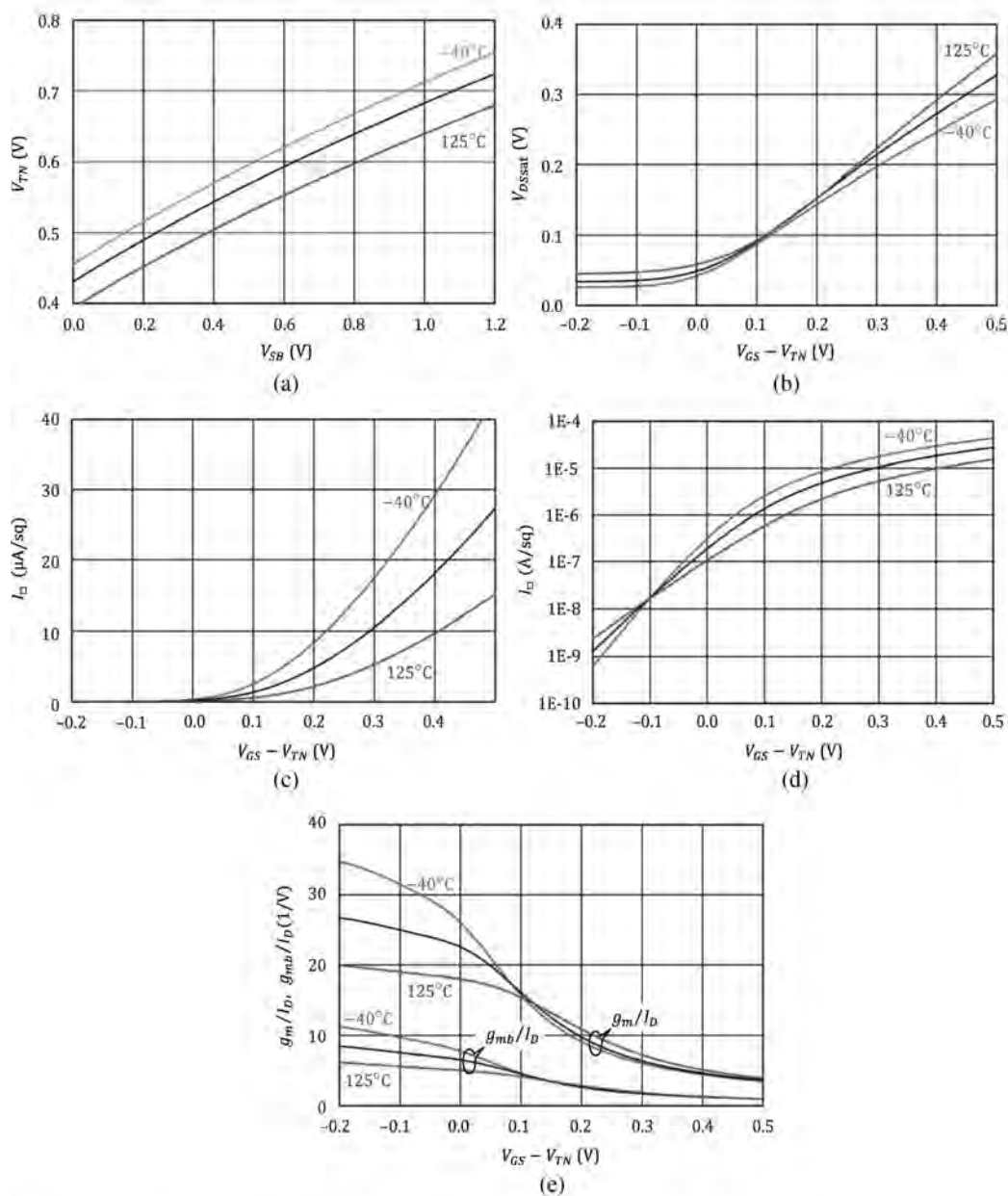


FIGURE 3.23 NMOS design-tool plots for $T = -40^\circ\text{C}$, $T = 27^\circ\text{C}$, and $T = 125^\circ\text{C}$. (a) Threshold voltage for $L_{DR} = 0.8 \mu\text{m}$. (b) Drain-source saturation voltage. (c) Sheet current on a linear axis. (d) Sheet current on a logarithmic axis. (e) Transconductance efficiencies.

3.6 Biasing and Sizing a MOSFET with Design Tools

基于设计工具设定MOS晶体管偏置和尺寸

As pointed out in Subsection 3.2.2, the main objective of MOSFET design is to determine the channel dimensions and bias conditions for a given set of performance metrics. In this section, we present a design-by-simulation methodology for achieving this objective. It relies on the design tools presented in Sections 3.3 and 3.4 for a device operating in saturation and enables the designer to determine W_{DR} , L_{DR} , and V_{GS} . Neither V_{SB} nor V_{DS} is an outcome of this methodology. Instead, these variables are assumed to be imposed either by topological constraints, or by the gate-source biases of adjacent devices, or by circuit specifications. The methodology is illustrated in Fig. 3.24 with a chart, where V_T stands for V_{TN} or V_{TP} depending on the type of device to be designed. What follows is a list describing the features of this chart.

- The chart contains 19 nodes, each being identified by a variable. A variable is classified either as a performance metric (I_D , V_{DSsat} , r_{ds} , g_m , g_{mb}) or as an externally imposed bias voltage (V_{SB} , V_{DS}) or as an intermediate variable (V_T , $V_{GS} - V_T$, g_m/I_D , g_{mb}/I_D , $I_{\square}$, W/L , W , L , V_A) or as a design variable (W_{DR} , L_{DR} , and V_{GS}).
- Each circle represents an algebraic relation among the variables attached, for example, $V_A = r_{ds}I_D$, $W/L = I_D/I_{\square}$, $I_D = g_m/(g_m/I_D)$, or $V_{GS} = (V_{GS} - V_T) + V_T$.
- Each box containing figure references represents a graphical design tool developed in Section 3.3 for NMOS (upper reference) or in Section 3.4 for PMOS

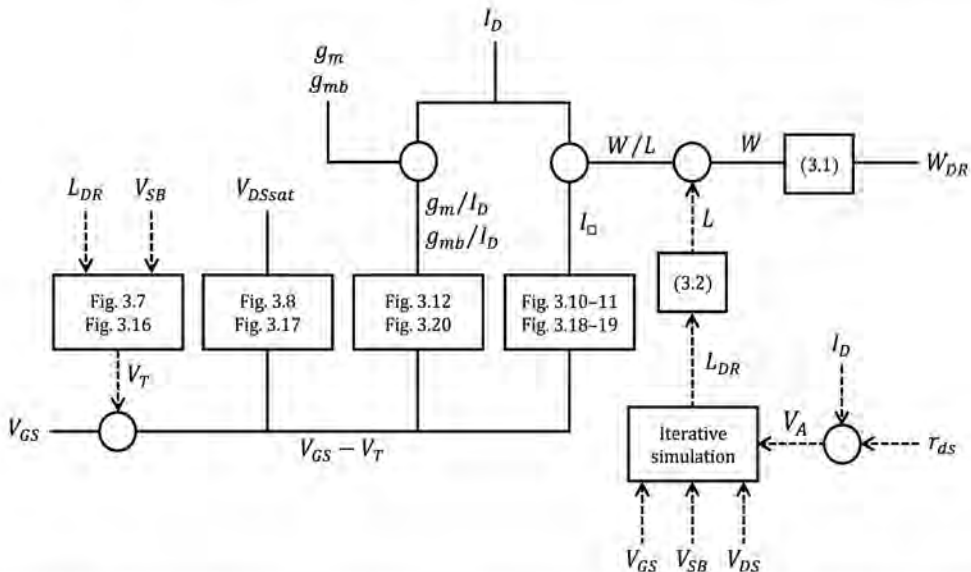


FIGURE 3.24 Flowchart of relations in a saturated MOSFET among device performance metrics (I_D , V_{DSsat} , r_{ds} , g_m , g_{mb}), externally imposed bias voltages (V_{SB} , V_{DS}), intermediate variables (V_T , $V_{GS} - V_T$, g_m/I_D , g_{mb}/I_D , $I_{\square}$, W/L , W , L , V_A), and design variables (W_{DR} , L_{DR} , and V_{GS}). Circles represent algebraic relations and boxes represent graphical design-tools and equations.

(lower reference). As stated previously, these tools are common to all devices of the same model deck.

- The box marked “iterative simulation” represents the relation between V_A and its variables. As discussed in Subsection 3.3.6, this relation involves multiple variables and cannot be generalized to all devices of the same model deck. An individual simulation is needed for each device whose biasing and sizing involves V_A .
- Boxes marked (3.1) and (3.2) represent the equations numbered accordingly.
- A solid line connecting a node to a circle or to a box defines a bidirectional path. The variable associated with the node may be moved into or out of the circle or box along such a line. For example, you may feed g_m/I_D into the box of Fig. 3.12 or Fig. 3.20, and thus determine the corresponding $V_{GS} - V_T$, or may do the opposite.
- A dashed line connecting a node to a circle or a box is unidirectional. The variable associated with the node should be moved only in the direction indicated by the arrowhead. For example, we may feed L_{DR} and V_{SB} into the box of Fig. 3.7 or Fig. 3.16 for the purpose of determining V_T , but should not attempt to determine V_{SB} or L_{DR} from V_T .

Let us now use the chart to determine the viable paths and steps leading from performance metrics to design variables W_{DR} , L_{DR} , and V_{GS} . As pointed out above, we assume prior knowledge of V_{DS} and V_{SB} . The chart indicates three well-constrained cases in which all three outcomes can be uniquely determined from specified performance metrics. These cases and their design flows are explained next.

1. V_{DSsat} , r_{ds} , and either g_m or g_{mb} are specified. Determine $V_{GS} - V_T$ from the design tool of Fig. 3.8 or Fig. 3.17, and translate the result into g_m/I_D (or g_{mb}/I_D) through the design tool of Fig. 3.12 or Fig. 3.20. Since g_m (or g_{mb}) is available as a specification item, calculate I_D . Next, determine $I_{\square}$ from the design tool of Figs. 3.10 to 3.11 or Figs. 3.18 to 3.19 and calculate W/L from I_D and $I_{\square}$. Also calculate V_A from I_D and r_{ds} . It is now possible to determine L_{DR} and V_{GS} by simulation. Set the simulation environment with (a) the specified values of V_{SB} and V_{DS} , (b) approximately $20W_{DR(min)}$ for W_{DR} , and (c) an arbitrary value for L_{DR} and the corresponding value for V_{GS} . The latter is calculated from $V_{GS} - V_T$ and V_T , where V_T is determined from the design tool of Fig. 3.7 or Fig. 3.16 for the specified value of V_{SB} and the arbitrarily selected value of L_{DR} . Run an operation-point analysis and read the value of V_A . If the latter turns out to be smaller than its target value calculated previously, then, increment L_{DR} , update the corresponding value of V_{GS} , and rerun the simulation. Otherwise, decrement L_{DR} , update V_{GS} , and rerun the simulation. You don't need to readjust W_{DR} in these iterations until the target value of V_A has been reached. Once the final values of L_{DR} and V_{GS} have thus been determined, calculate L from (3.2), and use the outcome together with W/L for determining W . Finally, calculate W_{DR} from (3.1).
2. I_D and r_{ds} and either g_m or g_{mb} are specified. Calculate g_m/I_D (or g_{mb}/I_D) from g_m (g_{mb}) and I_D . Determine firstly $V_{GS} - V_T$ from the design tool of Fig. 3.12 or Fig. 3.20, and subsequently $I_{\square}$ from the design tool of Figs. 3.10 to 3.11 or

Figs. 3.18 to 3.19. Now, calculate W/L from I_D and $I_{\square}$, and V_A from I_D and r_{ds} . Refer to the first case for iterative simulation and subsequent calculations.

3. I_D and r_{ds} and V_{DSsat} are specified. Determine $V_{GS} - V_T$ from the design tool of Fig. 3.8 or Fig. 3.17, and $I_{\square}$ from the design tool of Figs. 3.10 to 3.11 or Figs. 3.18 to 3.19. Next, calculate W/L from I_D and $I_{\square}$. Refer to the first case for iterative simulation and subsequent calculations.

Also note that specifying all three of I_D , V_{DSsat} , and g_m (or g_{mb}) simultaneously would result in an overconstrained case in which no unique value can be found for $V_{GS} - V_T$. At most two out of these three metrics can be specified without overconstraining the design. There may be underconstrained cases as well. Any such case provides the designer with an opportunity to impose his/her own specification(s) in accordance with good-design principles. Also note that there are many design scenarios in which one or more of W_{DR} , L_{DR} , and V_{GS} are topologically imposed or directly specified rather than being left as free-design variables. You can still use the chart of Fig. 3.24 in all these cases as long as you stick to the rules of use as itemized above. Several examples highlighting the universality of the chart follow next.

Example 3.6 An NMOS device is specified by $V_{SB} = 0$, $V_{DSsat} = 0.2$ V, $g_m = 2.1$ mA/V, and $r_{ds} = 47$ k Ω at $V_{DS} = 0.7$ V. Design and verify.

Work This is an example of the first well-constrained case described above. Design flow is as follows.

- $V_{DSsat} = 0.2$ V $\rightarrow$ Fig. 3.8 $\Rightarrow V_{GS} - V_{TN} = 0.275$ V.
- $V_{GS} - V_{TN} = 0.275$ V $\rightarrow$ Fig. 3.12 $\Rightarrow g_m/I_D = 7$ V $^{-1}$.
- $g_m/I_D = 7$ V $^{-1}$ and $g_m = 2.1$ mA/V $\Rightarrow I_D = 300$ μ A.
- $V_{GS} - V_{TN} = 0.275$ V $\rightarrow$ Fig. 3.10 $\Rightarrow I_{\square} = 9$ μ A/sq.
- $I_D = 300$ μ A and $I_{\square} = 9$ μ A/sq $\Rightarrow W/L = 33.3$.
- $I_D = 300$ μ A and $r_{ds} = 47$ k Ω $\Rightarrow V_A = 14.1$ V.
- Simulation: Input file is as follows:

```
filename: ex3-6.sp
.incl bsim.sp
m 2 1 0 0 cmosn w=6u l=0.4u
vgs 1 0 0.745
vds 2 0 0.7
.op
.end
```

Note that we arbitrarily set the drawn width to $W_{DR} = 20W_{DR(min)} = 6$ μ m in all iterations. The initial value of the channel length is also arbitrarily set to 0.4 μ m, for which $V_{TN} = 0.47$ V is indicated by Fig. 3.7. Since $V_{GS} - V_{TN}$ equals 0.275 V, V_{GS} is set to 0.745 V in this initial simulation. After each iteration, V_A is printed with the command

```
print @m[id]/@m[gds]
```

Channel length (μm)	0.4	0.6	0.7	0.8
Gate-source voltage (V)	0.745	0.722	0.715	0.705
Early voltage (V)	9.95	12.99	14.38	15.70

TABLE 3.3 Early Voltage versus Channel Length in Example 3.6

The result is shown in Table 3.3.

Obviously, $L_{DR} = 0.7 \mu\text{m}$ satisfies the specified V_A .

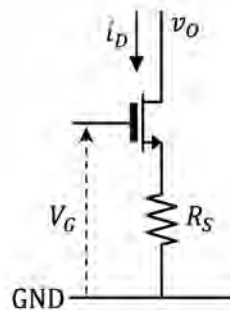
- $L_{DR} = 0.7 \mu\text{m} \rightarrow (3.2) \Rightarrow L = 0.66 \mu\text{m}$.
- $L = 0.66 \mu\text{m}$ and $W/L = 33.3 \Rightarrow W = 22.2 \mu\text{m}$.
- $W = 22.2 \mu\text{m} \rightarrow (3.1) \Rightarrow W_{DR} = 22.2 \mu\text{m}$. Actually, this step is redundant because it was obvious from the large effective width found in the preceding step that the relative difference between W and W_{DR} would be negligible.
- As always done in sizing a device, we need to verify the outcome of the design. We now rerun the input file after redefining the MOSFET and vgs element lines as follows:

```
m 2 1 0 0 cmosn w=22.2u l=0.7u
vgs 1 0 0.715
```

Verification results are as follows: $V_{DSsat} = 0.197 \text{ V}$, $g_m = 2.06 \text{ mA/V}$, $r_{ds} = 48.6 \text{ k}\Omega$, and $I_D = 294 \mu\text{A}$. ▲

Example 3.7 The circuit schematic shown in Fig. 3.25 belongs to a current sink, which is how we name an independent-current source if it directs the current into V_{SS} or ground. It remains functional as long as the MOSFET operates in saturation. The minimum output voltage that still satisfies this condition is called compliance limit, and is described by

$$V_{O(\min)} = V_{DSsat} + R_S I_D$$

**FIGURE 3.25** Current-sink application of Example 3.7.

Output resistance is yet another performance metric of a current sink. For the present sink, it is described by

$$r_o \doteq \left(\frac{\partial i_D}{\partial v_O} \right)^{-1},$$

which is related to MOSFET and resistor parameters through

$$r_o = r_{ds} + (g_m + g_{mb})r_{ds}R_S$$

as the reader will appreciate after studying Section 3.7.

Assume $R_S = 20 \text{ k}\Omega$, and consider an application specified by a compliance-limit $V_{O(\min)} = 0.3 \text{ V}$ in addition to a nominal current $I_D = 10 \text{ }\mu\text{A}$ and an output resistance $r_o = 5 \text{ M}\Omega$ at $V_O = 0.9 \text{ V}$. Size the device and determine the necessary gate bias V_G with respect to ground.

Work Design flow is as follows.

- For the specified values of $V_{O(\min)}$, R_S , and I_D , the expression presented above for the compliance-limit yields $V_{DSSat} = 100 \text{ mV}$, for which, Fig. 3.8 indicates $V_{GS} - V_{TN} = 0.115 \text{ V}$.
- $V_{GS} - V_{TN} = 0.115 \text{ V} \rightarrow \text{Fig. 3.10} \Rightarrow I_{\square} = 1.75 \text{ }\mu\text{A/sq}$.
- $I_D = 10 \text{ }\mu\text{A}$ and $I_{\square} = 1.75 \text{ }\mu\text{A/sq} \Rightarrow W/L = 5.71$.
- Multiply both sides of the output-resistance expression by I_D :

$$r_o I_D = r_{ds} I_D + \frac{g_m + g_{mb}}{I_D} r_{ds} I_D \times R_S I_D.$$

Since $r_{ds} I_D$ equals the Early voltage V_A , and $R_S I_D$ equals the source-bulk voltage V_{SB} , this equation can be rewritten as follows:

$$r_o I_D = V_A \left(1 + \frac{g_m + g_{mb}}{I_D} V_{SB} \right),$$

where $r_o I_D = 50 \text{ V}$, $V_{SB} = 0.2 \text{ V}$, and $(g_m + g_{mb})/I_D = 19 \text{ V}^{-1}$ as determined from Fig. 3.12 for $V_{GS} - V_{TN} = 0.115 \text{ V}$. Hence, $V_A = 10.4 \text{ V}$.

- We are now ready for iterative simulation. Before running a simulation, however, we first determine V_{TN} from the design tool of Fig. 3.7 for (a) $V_{SB} = 0.2 \text{ V}$, and (b) an arbitrary estimate of L_{DR} . We then add this V_{TN} to $V_{GS} - V_{TN} = 0.115 \text{ V}$

to obtain V_{GS} . Finally, adding $V_{SB} = 0.2$ V to V_{GS} will yield the voltage V_G to be applied between gate and ground. Input file is as follows:

```
filename: ex3-7a.sp
.incl bsim.sp
m 2 1 3 0 cmosn w=6u l=0.4u
vd 2 0 0.9
vg 1 0 0.838
vs 3 0 0.2
.op
.end
```

Iterative simulation eventually indicates an Early voltage of $V_A = 10.65$ V for $L_{DR} = 0.8$ μm .

- $L_{DR} = 0.8$ $\mu\text{m} \rightarrow (3.2) \Rightarrow L = 0.76$ μm .
- $L = 0.76$ μm and $W/L = 5.71 \Rightarrow W = 4.37$ $\mu\text{m} \Rightarrow W_{DR} = 4.4$ μm .
- Now that design is complete, we need to verify the specifications by simulating the entire circuit. Input file is as follows:

```
filename: ex3-7b.sp
.incl bsim.sp
m 2 1 3 0 cmosn w=4.4u l=0.8u
rs 3 0 20k
vo 2 0 0.9
vg 1 0 0.805
.tf v(2) vo
.end
```

Note that the verification of r_o is performed with a transfer-function analysis where vo is the input source, and $v(2)$ is the output variable. This type of analysis yields the following three pieces of information: (a) Small-signal transfer function from input source to output variable, which amounts to 1 in our case because vo and $v(2)$ are identical; (b) small-signal resistance loading the output port (called `output_impedance` in Spice terminology), which amounts to zero in our case because the independent voltage source vo shunts the output port of node 2 and ground; and (c) the small-signal resistance seen by the input source vo (called `input_impedance` in Spice terminology), which is exactly what the output resistance r_o of our current-source circuit is. For our present purposes, only this final piece of information of the transfer-function analysis is meaningful. Entering the command

```
print vo#input_impedance
```

yields $r_o = 5.02$ M Ω . Since Spice automatically runs an operation-point analysis prior to the transfer-function analysis, we also read $I_D = 10.1$ μA and $V_{DSsat} = 103$ mV simply by entering the command `show`.

The reader may have noticed that this example belongs to the third well-constrained case of design discussed prior to Example 3.6. ▲

Example 3.8 A PMOS device is specified by $V_{SB} = 0$ and $g_m = 160 \mu\text{A/V}$ in addition to $I_D = 16 \mu\text{A}$ and $r_{ds} = 2 \text{ M}\Omega$ at $V_{DS} = -0.9 \text{ V}$. Design and verify.

Work This is an example of the second well-constrained case discussed prior to Example 3.6. Design flow is as follows.

- $I_D = 16 \mu\text{A}$ and $g_m = 160 \mu\text{A/V} \Rightarrow g_m/I_D = 10 \text{ V}^{-1}$.
- $g_m/I_D = 10 \text{ V}^{-1} \rightarrow \text{Fig. 3.20} \Rightarrow V_{GS} - V_{TP} = -0.185 \text{ V}$.
- $V_{GS} - V_{TP} = -0.185 \text{ V} \rightarrow \text{Fig. 3.18} \Rightarrow I_{\square} = 1 \mu\text{A/sq}$.
- $I_D = 16 \mu\text{A}$ and $I_{\square} = 1 \mu\text{A/sq} \Rightarrow W/L = 16$.
- $I_D = 16 \mu\text{A}$ and $r_{ds} = 2 \text{ M}\Omega \Rightarrow V_A = 32 \text{ V}$.
- Simulation: Input file is as follows:

```
filename: ex3-8.sp
.incl bsim.sp
m 2 1 0 0 cmosp w=6u l=5.5u
vds 2 0 -0.9
vgs 1 0 -0.575
.op
.end
```

$L_{DR} = 5.5 \mu\text{m}$ satisfies the specification with $V_A = 32.39 \text{ V}$.

- $L_{DR} = 5.5 \mu\text{m} \rightarrow (3.2) \Rightarrow L = 5.44 \mu\text{m}$.
- $L = 5.44 \mu\text{m}$ and $W/L = 16 \Rightarrow W = 87.1 \mu\text{m}$.
- $W = 87.1 \mu\text{m} \rightarrow (3.1) \Rightarrow W_{DR} = 87.1 \mu\text{m}$.

Verification will indicate $g_m = 160 \mu\text{A/V}$, $I_D = 16.23 \mu\text{A}$, and $r_{ds} = 2.00 \text{ M}\Omega$. ▲

Example 3.9 An NMOS device is specified by $I_D = 150 \mu\text{A}$ and $g_m = 1.1 \text{ mA/V}$. Gate area WL should not exceed $65 \mu\text{m}^2$. Design for the largest possible r_{ds} .

Work Design flow is as follows:

- $I_D = 150 \mu\text{A}$ and $g_m = 1.1 \text{ mA/V} \Rightarrow g_m/I_D = 7.33 \text{ V}^{-1}$.
- $g_m/I_D = 7.33 \text{ V}^{-1} \rightarrow \text{Fig. 3.12} \Rightarrow V_{GS} - V_{TN} = 0.26 \text{ V}$.
- $V_{GS} - V_{TN} = 0.26 \text{ V} \rightarrow \text{Fig. 3.10} \Rightarrow I_{\square} = 7 \mu\text{A/sq}$.
- $I_D = 150 \mu\text{A}$ and $I_{\square} = 7 \mu\text{A/sq} \Rightarrow W/L = 21.4$.
- Since the largest possible r_{ds} is desired, L must be maximized. Therefore adopt the largest permissible gate area. Solve $WL = 65 \mu\text{m}^2$ together with $W/L = 21.4$.
 $\Rightarrow W = 37.3 \mu\text{m}$, $L = 1.74 \mu\text{m}$. ▲

3.7 Small-Signal Modeling and Circuit Analysis

小信号模型和电路分析

In Section 3.6, we developed a design-by-simulation technique for sizing and biasing a MOSFET for a given set of device performance metrics. Three of these performance met-

rics, g_m , g_{mb} , and r_{ds} are defined in terms of variations in port variables. The relationship between these three and open-loop performance metrics are derived with a *small-signal analysis* of the opamp. MOSFET modeling for dc small-signal analysis is the first subject of this section. Based on this model, we will develop a technique by which we can determine the dc small-signal behavior of an analog circuit by visual inspection of the electrical schematic only. Finally, we will discuss the parasitic capacitances of a MOSFET, and develop a high-frequency small-signal device model based on these parasitics.

3.7.1 MOSFET DC Small-Signal Model MOS晶体管直流小信号模型

MOSFET drain current i_D depends nonlinearly on port voltages v_{GS} , v_{DS} , and v_{SB} . We can represent the relation generally with

$$i_D = f(v_{GS}, v_{DS}, v_{SB}). \quad (3.53)$$

Now, suppose that these three port voltages vary around their respective bias voltages V_{GS} , V_{DS} , and V_{SB} , and variations are represented by v_{gs} , v_{ds} , and v_{sb} . We can expand (3.53) into Taylor series around the bias voltages and find the corresponding variation i_d in the drain current i_D around its bias value I_D . If we adopt small-signal approximation, expansion will involve only the first-order terms. Hence,

$$i_d \approx \left. \frac{\partial i_D}{\partial v_{GS}} \right|_{v_{GS}=V_{GS}} v_{gs} + \left. \frac{\partial i_D}{\partial v_{DS}} \right|_{v_{DS}=V_{DS}} v_{ds} + \left. \frac{\partial i_D}{\partial v_{SB}} \right|_{v_{SB}=V_{SB}} v_{sb}. \quad (3.54)$$

According to the definitions (3.5), (3.6), and (3.7), we can rewrite (3.54) as

$$i_d = g_m v_{gs} + \frac{1}{r_{ds}} v_{ds} - g_{mb} v_{sb} \quad (3.55)$$

for an NMOS device, and $i_d = -g_m v_{gs} - (1/r_{ds})v_{ds} + g_{mb}v_{sb}$ for a PMOS. Recall that these equations comply with the reference convention set for the MOSFET variables in Fig. 3.2; that is, i_D is positive if it flows into the drain in NMOS, or out of drain in PMOS. Now on, we transpose the reference direction of i_d for PMOS, so that it will be positive when it flows into the drain exactly like the case of NMOS. This enables us to use (3.55) to describe the small-signal drain current of both types of devices. Do not forget that we transpose the reference direction in PMOS only for i_d and not for i_D . In other words, we adopt $i_D = I_D - i_d$ for PMOS while retaining $i_D = I_D + i_d$ for NMOS. Another fact to bear in mind is that g_m , r_{ds} , and g_{mb} are characterized for the particular bias voltages around which the device operates.

Equation (3.55) relates the dc small-signal drain current to dc small-signal port voltages. It is accurate only for small dc signals. In dc small-signal analysis of a circuit, we can represent a MOSFET with a *small-signal equivalent circuit* based on this equation. The equivalent circuit is shown in Fig. 3.26.

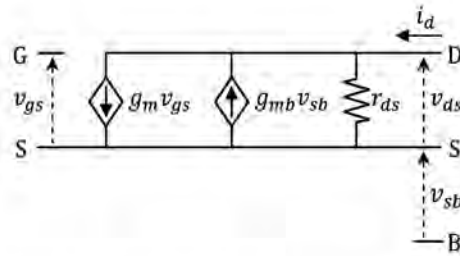


FIGURE 3.26 MOSFET dc small-signal equivalent circuit.

The reason why we qualify the equivalent circuit of Fig. 3.26 for dc excitation only is that the capacitive effects associated with MOSFET operation are not taken into account in developing (3.55). Nevertheless, we rely on this equivalent circuit also in ac analyses as long as the frequency of operation is sufficiently low to make capacitive currents negligible. It is therefore more descriptive to name it as *dc and low-frequency equivalent circuit*.

3.7.2 DC Small-Signal Circuit Analysis 直流小信号电路分析

Formally, the dc small-signal analysis of an analog circuit is performed on an equivalent circuit obtained by applying the following procedure:

- Step 1* Replace each capacitance with an open circuit since a capacitive impedance is infinite for dc.
- Step 2* Replace each inductance with a short circuit since an inductive impedance is zero for dc.
- Step 3* Replace each dc bias or supply source with an open circuit in the case of a current source, and by a short circuit in the case of a voltage source.
- Step 4* Replace each MOSFET with the equivalent circuit given in Fig. 3.26.

Let us now illustrate the application of this formal procedure with two simple examples.

Example 3.10 Formulate the dc voltage gain of the amplifier of Example 3.5.

Work The amplifier schematic given in Fig. 3.22(a) is translated into a dc small-signal equivalent circuit in Fig. 3.27(a) where gray-shaded boxes represent the equivalent circuits of M1 and M2. Referring also to Fig. 3.26, notice the following features of the equivalent circuit:

- Both MOSFETs operate with a source/bulk short, which implies $v_{sb} = 0$ for both devices. The dependent current-source $g_{mb}v_{sb}$ thus becoming an open circuit is removed from the equivalent circuits of both MOSFETs.
- Short-circuiting the dc voltage sources V_{DD} and V_B implies grounded source and gate for M2. Since v_{gs} is therefore zero for M2, the dependent current-source $g_m v_{gs}$ is also replaced by an open circuit in M2.

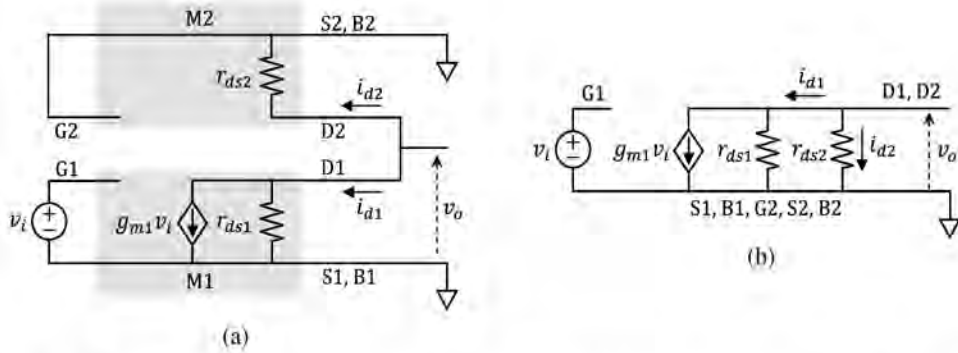


FIGURE 3.27 (a) Detailed dc small-signal equivalent circuit of the common-source amplifier stage of Example 3.10. (b) Usual depiction of the equivalent circuit.

- Input signal v_i is applied between the gate and source of M1. Therefore, the dependent current-source $g_m v_{gs}$ equals $g_{m1} v_i$ in M1.
- Output signal v_o is defined between the common drain of both devices and ground.
- The reference direction of drain signal currents i_{d1} and i_{d2} is directed from drain toward the inside of both devices. Notice $i_{d2} = -i_{d1}$.

Figure 3.27(b) depicts a more usual form of the equivalent circuit, which indicates

$$v_o = -g_{m1} (r_{ds1} \parallel r_{ds2}) v_i.$$

Therefore, voltage gain is formulated as

$$A_v \triangleq \frac{v_o}{v_i} = -g_{m1} (r_{ds1} \parallel r_{ds2}) = -g_{m1} \frac{r_{ds1} r_{ds2}}{r_{ds1} + r_{ds2}}. \quad \blacktriangle \quad (3.56)$$

Example 3.11 Shown in Fig. 3.28(a) is a diode-connected PMOS device operating with a resistor R_S between its source and power-supply rail V_{DD} . Determine the small-signal resistance r seen between the common gate/drain node and signal ground.

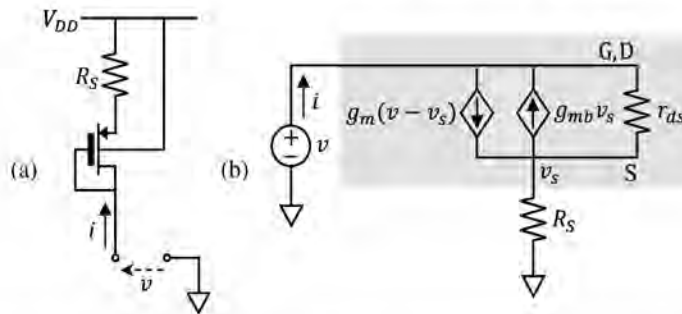


FIGURE 3.28 (a) A diode-connected PMOS with a resistor at its source. (b) Small-signal equivalent circuit for determining the resistance seen into the common gate/drain node.

Work The small-signal resistance is defined by $r \triangleq v/i$, where v is a small-signal voltage applied to the common gate/drain node with respect to ground, and i is the small-signal current flowing into the MOSFET. The small-signal equivalent of the circuit is shown in Fig. 3.28(b), where the MOSFET equivalent circuit is shaded gray, and the dc voltage source V_{DD} is grounded. Note that i is the sum of three branch currents as expressed by

$$i = g_m(v - v_s) - g_{mb}v_s + \frac{v - v_s}{r_{ds}},$$

and v_s is related to i by $v_s = iR_s$. Solving for r , we obtain

$$r = \frac{r_{ds}}{1 + r_{ds}g_m} (1 + g_{mb}R_s) + R_s.$$

Diode-connected MOSFET finds widespread use in analog circuits. In most cases, it does not contain a source resistor; instead, its source is tied directly to a power-supply rail or ground. In such cases, the expression of small-signal resistance is reduced to

$$r = \frac{r_{ds}}{1 + r_{ds}g_m},$$

$r_{ds}g_m$ is usually much larger than unity, which enables us to simplify this equation into

$$r = \frac{1}{g_m}. \quad \blacktriangle$$

In Examples 3.10 and 3.11, we determined the required expressions rapidly and painlessly because the circuit topologies involved were very simple. In more complex topologies, the formal method of circuit analysis with MOSFET equivalent circuits may prove highly time consuming and error prone. Fortunately, the dc small-signal behavior of a circuit can be formulated by inspection without resorting to formal circuit analysis. The methodology of inspection is based on a number of generic rules by which the voltage transfer function between two terminals of a MOSFET is constructed. Note that a MOSFET whose bulk terminal is tied to its source terminal or to ground or to a dc bias-voltage source can transfer a dc signal along the following four paths: (a) from gate to source, (b) from gate to drain, (c) from source to drain, and (d) from drain to source. In rare cases, bulk may be tied neither to the source terminal nor to a bias source or ground but to an independent signal source. In such cases, two additional signal flow paths are created: (e) from bulk to source, and (f) from bulk to drain. Bearing in mind these possible directions of signal flow, let us consider the exemplar circuit shown in Fig. 3.29(a) for a first taste of analysis by inspection. This is a split-supply *differential amplifier* operating between a positive rail V_{DD} and a negative rail V_{SS} , and receiving its input signal v_i at the gate of M1 with respect to ground. All NMOS devices have their bulks at V_{SS} ,

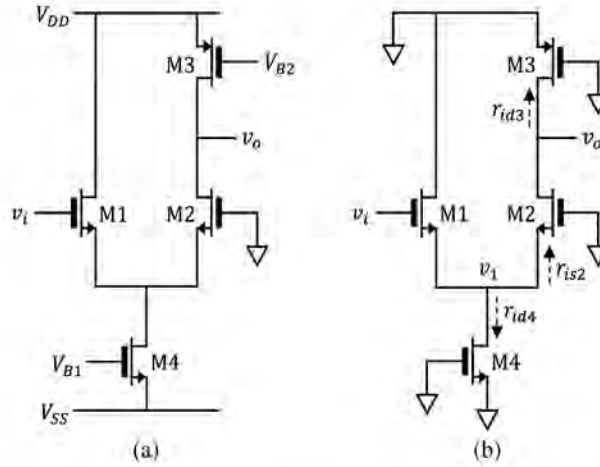


FIGURE 3.29 (a) A differential amplifier stage. (b) Equivalent circuit for analysis by inspection.

whereas the bulk of M3 is at V_{DD} . Output signal v_o is defined between the drain of M2 and ground. Suppose that our objective is to formulate the voltage gain

$$A_v \triangleq \frac{\hat{v}_o}{\hat{v}_i}.$$

Formally, we would sketch an equivalent circuit applying all four steps of the procedure described at the beginning of this subsection. In analysis by inspection, we apply the first three only. We do not replace devices with equivalent circuits; instead, we retain their symbols in the equivalent circuit. In the present example, the first two steps are redundant anyway because no capacitance or inductance is present in the exemplar circuit. Step 3 is applicable to supply sources V_{DD} and V_{SS} as well as bias sources V_{B1} and V_{B2} . After replacing these with short circuits, we arrive at the equivalent circuit shown in Fig. 3.29(b). Next, starting at the gate of M1, we identify the signal-flow path from input to output on this equivalent circuit. Signal generally can flow from gate to drain and source but, in the case of M1, it cannot reach the output through the drain because the latter is grounded. This is why we consider only its flow to the source, where its voltage is denoted by v_1 . M2 receives this signal at its own source terminal, and passes it to its drain as the output signal v_o . Notice in the present example that the signal flows from input to output through a single forward path without looping back. In practice we may come across circuits containing multiple forward paths and/or feedback loops. In such cases, we apply superposition and/or loop-cutting techniques to transform the equivalent circuit into multiple circuits each comprising a single forward path. How we do these transformations will be explained and exemplified later. Let us continue with our present example.

Having traced the path of signal flow, we now factorize the overall gain into individual gains of the devices located along the signal path:

$$A_v = \frac{v_o}{v_i} = \frac{v_o}{v_1} \frac{v_1}{v_i}.$$

Notice that each factor represents a *voltage gain* from an *input terminal* to an *output terminal* of the same MOSFET. The first is from source to drain in M2, and the second is from gate to source in M1. As we will soon discover, voltage gain from one terminal of a MOSFET to another terminal of the same MOSFET in a single forward path has a very general expression regardless of the topology of the circuit in which this MOSFET resides. In such an expression, the part of the circuit connecting the output terminal of the MOSFET is represented by an equivalent resistance between this terminal and circuit ground. As an example, consider M1 whose output terminal is the source. You can visualize the effective resistance connecting its source by inspecting Fig. 3.29(b). Obviously it is the parallel equivalent of two resistances: (a) r_{id4} , which is the drain input resistance of M4, and (b) r_{is2} , which is the source input resistance of M2. Moving to M2, whose output terminal on the signal path is drain, we observe the drain input resistance r_{id3} of M3 as the resistance loading its output. Fortunately, input resistances seen into the drain and source terminals of a MOSFET are also described by two very general expressions as long as the device resides on a single forward path.

We now derive generic voltage-gain and input-resistance expressions of analysis-by-inspection for four possible cases of signal entry through source, drain, gate, or bulk. Although each case will be illustrated with an NMOS device, the results will be perfectly valid also for a PMOS device thanks to the unified small-signal equivalent circuit presented in Subsection 3.7.1.

Source Drive

First consider a MOSFET whose bulk is tied to a supply or bias voltage or to ground, and suppose that it receives a dc or low-frequency small-signal voltage v_s at its source terminal with respect to circuit ground as shown in the small-signal equivalent circuit of Fig. 3.30(a). Note that the bulk terminal is shown grounded regardless of whether it is physically grounded or tied to a supply or bias source in the actual circuit. This is because all dc voltage sources are short-circuited in a small-signal analysis. Also notice the grounding of the gate terminal. Although gate is physically connected to some node of the circuit in which this MOSFET operates, it does not support a signal because (a) in a circuit of single forward signal path, only one of the MOSFET terminals can be an input terminal, which is assumed to be the source terminal in the present case, and (b) as explained previously, a dc or low-frequency signal applied to source cannot flow to gate. Finally, notice that the external circuit loading the drain terminal is represented by an equivalent resistance R_D placed between drain and ground. The output signal at the drain terminal, represented by v_d with respect to ground, is therefore the signal voltage across R_D .

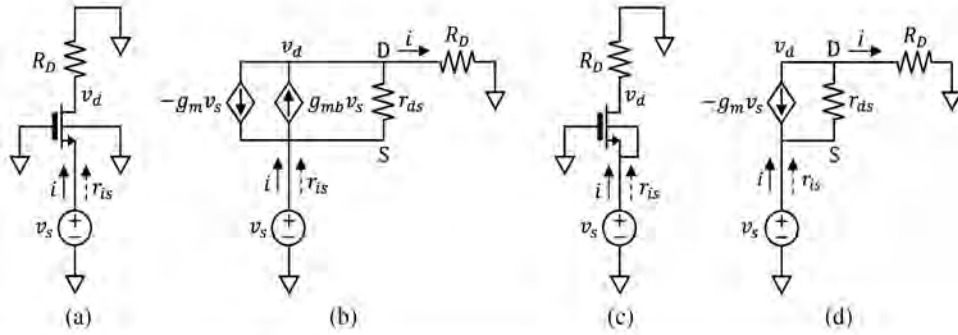


FIGURE 3.30 (a) A MOSFET operating without source-bulk short and receiving a small-signal input voltage at its source terminal. (b) Small-signal equivalent of the circuit shown in (a). (c) A MOSFET operating with source-bulk short and receiving a small-signal input signal at its source terminal. (d) Small-signal equivalent of the circuit shown in (c).

By replacing the MOSFET with its own equivalent circuit, we turn the circuit of Fig. 3.30(a) into what is shown in Fig. 3.30(b), where

$$i = (g_m + g_{mb})v_s + \frac{1}{r_{ds}}(v_s - v_d),$$

and

$$v_d = iR_D.$$

Using these two equations, the MOSFET input resistance r_{is} seen by the signal source between the source terminal and ground is formulated as

$$r_{is} \triangleq \frac{v_s}{i} = \frac{r_{ds} + R_D}{A_{\max}}, \quad (3.57)$$

where

$$A_{\max} = 1 + r_{ds}(g_m + g_{mb}) \approx r_{ds}(g_m + g_{mb}) \quad (3.58)$$

is the maximum voltage gain offered by a MOSFET, as will be proven soon. The approximation made in (3.58) is due to the fact that the product $r_{ds}(g_m + g_{mb})$ is typically around 100.

In order to formulate the gain from source to drain, we replace i in $r_{is} \triangleq v_s/i$ with $i = v_d/R_D$, and therefore obtain

$$\frac{v_d}{v_s} = \frac{R_D}{r_{is}}. \quad (3.59)$$

We can now prove that $A_{\max}$ is the maximum gain of a MOSFET. Just substitute (3.57) into (3.59) to obtain

$$\frac{v_d}{v_s} = \frac{R_D}{r_{ds} + R_D} A_{\max},$$

which obviously is an increasing function of R_D . At the limit $R_D \rightarrow \infty$, the gain is indeed maximized at $A_{\max}$.

Next, consider a MOSFET whose bulk terminal is tied to its source terminal, and suppose that the common source/bulk terminal receives a dc or low-frequency small signal $\bar{v}_s$ with respect to ground as shown in Fig. 3.30(c). The input resistance seen into the source/bulk terminal and the voltage gain from this common terminal to drain can be formulated by analyzing the small-signal equivalent circuit shown in Fig. 3.30(d). Note that the dependent current source $g_{mb}v_{sb}$ is missing because of the source-bulk short. Equivalent-circuit analysis yields exactly (3.57) for r_{is} and (3.59) for v_d/v_s . The maximum gain, however, is described by

$$A_{\max} = 1 + r_{ds}g_m \approx r_{ds}g_m \quad (3.60)$$

Drain Drive

Now we apply a dc small-signal voltage v_d to the drain of a MOSFET whose bulk is either grounded or tied to a supply or bias source. As shown in Fig. 3.31(a), the signal can flow only to source, which is assumed to be loaded by an external circuit of an equivalent resistance R_S . The corresponding small-signal equivalent circuit is shown in Fig. 3.22(b). Its analysis yields

$$r_{id} = r_{ds} + A_{\max}R_S \quad (3.61)$$

for drain input resistance. Note that $A_{\max}$ is still described by (3.58) in this case. The gain from drain to source is described by

$$\frac{v_s}{v_d} = \frac{R_S}{r_{id}} \quad (3.62)$$

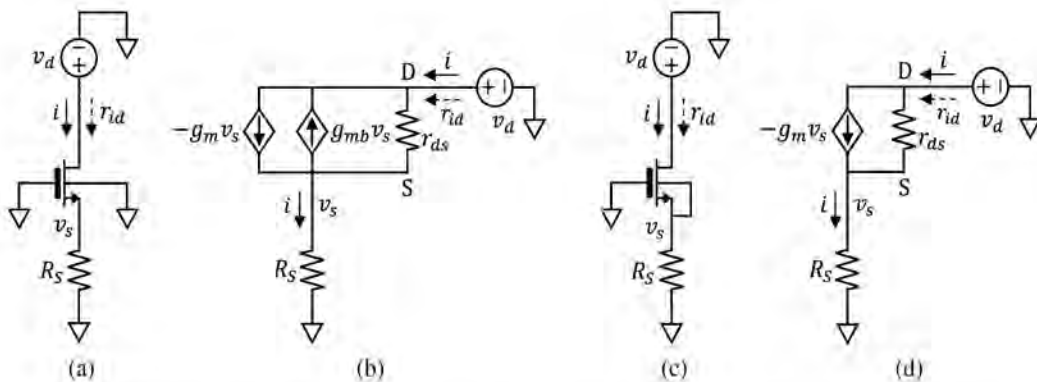


FIGURE 3.31 (a) A MOSFET operating without source-bulk short and receiving a small-signal input voltage at its drain terminal. (b) Small-signal equivalent of the circuit shown in (a). (c) A MOSFET operating with source-bulk short and receiving a small-signal input signal at its drain terminal. (d) Small-signal equivalent of the circuit shown in (c).

If the MOSFET receiving an input signal at the drain terminal operates with a source-bulk short as depicted in Fig. 3.31(c), then, small-signal analysis will be based on the equivalent circuit shown in Fig. 3.31(d). Analysis again yields (3.61) and (3.62) for drain input resistance and gain, respectively, but $A_{\max}$ is described by (3.60) in this case.

Gate Drive

Shown in Fig. 3.32(a) is a MOSFET whose bulk is either grounded or tied to a supply or bias source, and gate receives a small-signal input voltage v_g . Since a signal applied to the gate terminal can flow to drain and source, both of the latter two can be designated as output terminals. Let R_D and R_S represent the equivalent resistance of the external circuit loading the drain and source, respectively. The corresponding equivalent circuit is given in Fig. 3.32(b). Note that the input resistance seen by the signal source is infinite. Analyzing the circuit and doing some algebra, we arrive at the following generic gain expressions for signal flow from gate to drain, and from gate to source:

$$\frac{v_d}{v_g} = -\frac{r_{ds}g_m}{1 + r_{ds}(g_m + g_{mb})} \frac{R_D}{R_S + r_{is}} \approx -\frac{g_m}{g_m + g_{mb}} \frac{R_D}{R_S + r_{is}} \quad (3.63)$$

$$\frac{v_s}{v_g} = \frac{r_{ds}g_m}{1 + r_{ds}(g_m + g_{mb})} \frac{R_S}{R_S + r_{is}} \approx \frac{g_m}{g_m + g_{mb}} \frac{R_S}{R_S + r_{is}} \quad (3.64)$$

where approximations are again due to the fact that $r_{ds}(g_m + g_{mb})$ is much larger than unity.

The case of a MOSFET operating with a source-bulk short and receiving an input signal v_g at its gate is shown in Fig. 3.32(c). The corresponding small-signal equivalent circuit is given in Fig. 3.32(d). An analysis of the latter yields the following generic gain expressions:

$$\frac{v_d}{v_g} = -\frac{r_{ds}g_m}{1 + r_{ds}g_m} \frac{R_D}{R_S + r_{is}} \approx -\frac{R_D}{R_S + r_{is}} \quad (3.65)$$

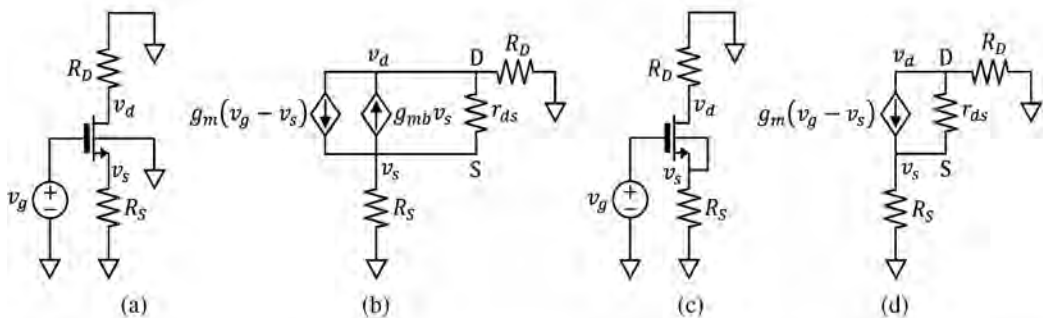


FIGURE 3.32 (a) A MOSFET operating without source-bulk short and receiving a small-signal input voltage at its gate terminal. (b) Small-signal equivalent of the circuit shown in (a). (c) A MOSFET operating with source-bulk short and receiving a small-signal input signal at its gate terminal. (d) Small-signal equivalent of the circuit shown in (c).

and

$$\frac{v_s}{v_g} = \frac{r_{ds}g_m}{1 + r_{ds}g_m} \frac{R_S}{R_S + r_{is}} \approx \frac{R_S}{R_S + r_{is}}. \quad (3.66)$$

Bulk Drive

As mentioned previously, in some rare cases, a MOSFET may receive an input signal only at its bulk terminal as shown in Fig. 3.33(a). This signal can flow to source or drain. The equivalent circuit appropriate for this case is shown in Fig. 3.33(b). Its analysis yields the following gain expressions:

$$\frac{v_d}{v_b} = -\frac{r_{ds}g_{mb}}{1 + r_{ds}(g_m + g_{mb})} \frac{R_D}{R_S + r_{is}} \approx -\frac{g_{mb}}{g_m + g_{mb}} \frac{R_D}{R_S + r_{is}}, \quad (3.67)$$

$$\frac{v_s}{v_b} = \frac{r_{ds}g_{mb}}{1 + r_{ds}(g_m + g_{mb})} \frac{R_S}{R_S + r_{is}} \approx \frac{g_{mb}}{g_m + g_{mb}} \frac{R_S}{R_S + r_{is}}, \quad (3.68)$$

where approximations are again due to the fact that $r_{ds}(g_m + g_{mb})$ is much larger than unity. Notice the similarity between these two expressions and (3.63) and (3.64) of gate drive. The only difference is the transposed roles of g_m and g_{mb} .

Rules of Analysis by Inspection

Analysis by inspection is based on two input-resistance expressions for r_{is} and r_{id} , and six gain expressions for v_d/v_s , v_s/v_d , v_d/v_g , v_s/v_g , v_d/v_b , and v_s/v_b . These expressions can be interpreted in terms of a few common rules which are much easier to remember than the expressions themselves. We now present these rules.

1. *Input Resistances* The four resistances seen when viewed from outside or inside a MOSFET through source or drain terminals are illustrated in Fig. 3.34. The main rule to remember when determining these resistances by inspection is that the source terminal magnifies a resistance by A_{max} when the resistance is viewed from inside the device, and demagnifies it by the same factor when the

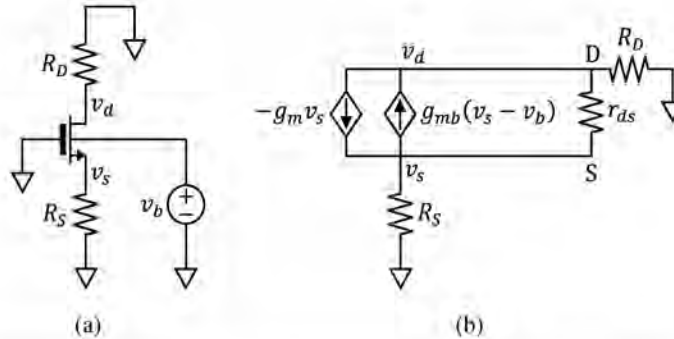


FIGURE 3.33 (a) A MOSFET operating without source-bulk short and receiving a small-signal input voltage at its bulk terminal. (b) Small-signal equivalent of the circuit of (a).

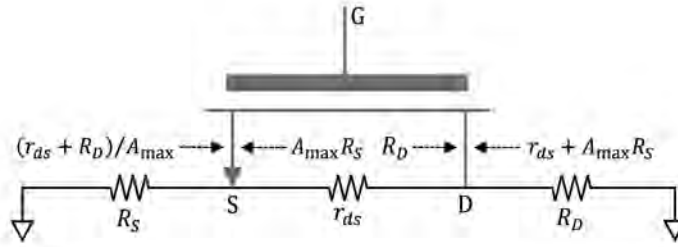


FIGURE 3.34 Resistance magnification and demagnification at the source terminal of a MOSFET.

resistance is viewed from outside. The drain terminal has no such property; it does not magnify or demagnify the value of a resistance regardless of the direction in which you view the resistance. Let us apply these rules firstly for constructing the expression of r_{id} . Looking into the device from outside the drain, we immediately see r_{ds} between drain and source without any magnification or demagnification. There exists also an external resistance R_S lying in series with r_{ds} between the source and ground but since we are viewing it from inside out through the source terminal, what we see is $A_{\max} R_S$. Since the two resistances are in series, the equivalent resistance equals their sum, hence (3.61). If the source terminal is grounded ($R_S = 0$), we obviously will see only r_{ds} for r_{id} .

Let us now apply the rule to r_{is} . Since drain does not modify the view of a resistance, the total resistance seen between the inside of the source terminal and ground through drain equals $r_{ds} + R_D$. However, r_{is} is the resistance seen from outside the source. Since source demagnifies by $A_{\max}$ when viewed from outside, we indeed end up with (3.57) for r_{is} . Also note that if the drain is held at signal ground ($R_D = 0$), r_{is} will consist only of $r_{ds}/A_{\max}$, which equals $1/(g_m + g_{mb})$ or $1/g_m$ depending on whether the bulk is at signal ground or tied to the source, respectively.

2. **Maximum Gain** Maximum gain $A_{\max}$ equals the product $r_{ds}(g_m + g_{mb})$ in the case of a bulk tied to ground or to a dc voltage source. In a device operating with a source-bulk short, maximum gain equals $r_{ds}g_m$ only.
3. **Source-to-Drain Gain and Drain-to-Source Gain** These gain expressions are constructed by dividing the external resistance between the output terminal and ground (R_D for drain output and R_S for source output) by input resistance; hence, (3.59) for source-to-drain gain and (3.62) for drain-to-source gain.

Considering (3.61) and (3.62) together, we conclude that the gain v_s/v_d from drain to source equals $1/A_{\max}$ for $R_S \rightarrow \infty$ and vanishes for $R_S = 0$. This is why we never use this signal path when we intend to obtain voltage gain from a MOSFET.

4. **Gate-to-Drain Gain and Gate-to-Source Gain** As indicated by (3.63) and (3.64), each of these two gain expressions is generally a product of two factors. One of these factors is the ratio of the external resistance between the output terminal and ground (R_D for drain output and R_S for source output) to the total resistance $R_S + r_{is}$ associated with the source terminal. The second factor is the ratio $g_m/(g_m + g_{mb})$, which equals unity in the case of a device operating with a

source-bulk short as in (3.65) and (3.66). Also note that the gate-to-drain gain expression contains a minus sign, which indicates that v_d varies in the opposite direction of v_g .

5. *Bulk-to-Drain Gain and Bulk-to-Source Gain* Except for the transposal of g_m and g_{mb} , these gain expressions are similar to the gate-to-drain and gate-to-source counterparts.

Example 3.12 Determine by inspection the small-signal voltage gain v_o/v_i of the common-source amplifier stage shown in Fig. 3.22(a), and compare the result with (3.56).

Work Gain v_o/v_i is from gate to drain in M1. Also note that none of the devices has an external source resistance R_S , while both operate with source-bulk shorts. After grounding V_{DD} and V_B , we obtain the equivalent circuit shown in Fig. 3.35.

Adopting gate-drive expressions, we analyze the circuit as follows.

- (3.65) $\Rightarrow \frac{v_o}{v_i} = -\frac{R_{D1}}{r_{is1}}$
- Figure 3.35 $\Rightarrow R_{D1} = r_{id2}$, and (3.61) $\Rightarrow r_{id2} = r_{ds2}$
- (3.57) $\Rightarrow r_{is1} = \frac{r_{ds1} + r_{ds2}}{A_{max1}}$
- (3.60) $\Rightarrow A_{max1} = r_{ds1}g_{m1}$
- Combining above equations, we arrive at

$$\frac{v_o}{v_i} = -g_{m1} \frac{r_{ds1}r_{ds2}}{r_{ds1} + r_{ds2}}$$

which is identical to (3.56). ▲

Example 3.13 Determine by inspection the small-signal voltage gain v_o/v_i of the differential amplifier shown in Fig. 3.29. Assume (a) a twin-well CMOS technology where chip substrate is the common bulk of all NMOS devices, and (b) identical M1 and M2 whose gate and bulk transconductances are denoted by g_m and g_{mb} .

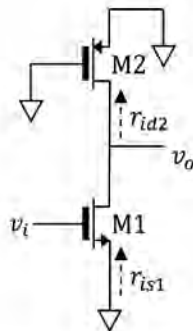


FIGURE 3.35 Small-signal equivalent of the circuit of Fig. 3.22(a).

Work Since the substrate has to be tied to V_{SS} , all NMOS devices operate with a common bulk voltage at V_{SS} . This implies a source/bulk short only for M4 among the three NMOS devices. The PMOS device M3 also operates with a source/bulk short because its source is already at the highest available voltage V_{DD} . Relying on the small-signal equivalent circuit given in Fig. 3.29(b), we identify a signal path from the gate of M1 to the source of M1, and from the source of M2 to the drain of M2. Therefore, the circuit has a single forward path from input to output. We can directly apply the appropriate gain and input-resistance expressions derived in this section, and arrive at

$$\frac{v_o}{v_i} = \frac{g_m}{g_m + g_{mb}} \underbrace{\frac{r_{ds4} \parallel r_{is2}}{(r_{ds4} \parallel r_{is2}) + r_{is1}}}_{v_1/v_i} \underbrace{\frac{r_{ds3}}{r_{is2}}}_{v_o/v_1}, \quad (3.69)$$

where

$$r_{is1} = \frac{1}{g_m + g_{mb}},$$

and

$$r_{is2} = \frac{r_{ds3} + r_{ds2}}{r_{ds2}(g_m + g_{mb})}.$$

It is obvious that the final gain expression obtained from these three equations will be pretty complicated. Fortunately, we can simplify it with an acceptable approximation based on the fact that, in most cases of small-signal analysis, we may assume $1/(g_m + g_{mb})$ to be negligibly small in comparison with an r_{ds} even if the two belong to two different devices. To that end, first, consider the expression of r_{is2} . Since we expect r_{ds2} and r_{ds3} to be on the same order of magnitude, r_{is2} should be comparable to $1/(g_m + g_{mb})$. Since the latter is expected to be much smaller than r_{ds4} , the parallel equivalent $r_{ds4} \parallel r_{is2}$ appearing in (3.69) equals r_{is2} to a good approximation. Also considering the expressions of r_{is1} and r_{is2} , we simplify (3.69) into

$$\frac{v_o}{v_i} \approx g_m \frac{r_{ds3} r_{ds2}}{r_{ds3} + 2r_{ds2}}. \quad \blacktriangle$$

Example 3.14 Shown in Fig. 3.36(a) is the electrical schematic of the split-supply version of a very popular single-stage operational transconductance amplifier called *symmetrical OTA*. I_B is a dc bias-current source assumed to be ideal in this example. M1 and M2 have identical dimensions, and their gates are biased at ground. The diode-connected M3 and M4 also have identical dimensions. Furthermore, M5 and M7 have the same channel length but not necessarily the same channel width. Same is valid for M6 and M8. Needless to say, the wells of all four PMOS devices are tied to V_{DD} . We first analyze the bias condition of the circuit in some detail, and then perform a small-signal analysis for the purpose of determining the voltage gain v_o/v_i .

Work First, notice from Fig. 3.36(a) that M1 and M2 operate with the same gate-source bias voltage V_{GS} because (a) their gates are commonly tied to ground, and (b) their source terminals are physically interconnected. Having the same V_{SB} and the same L_{DR} , they are identical also in V_{TN} . Therefore, both operate with the same overdrive $V_{GS} - V_{TN}$, hence the same $I_{\square}$ (see Fig. 3.24), which, in turn, implies identical bias currents ($I_B/2$

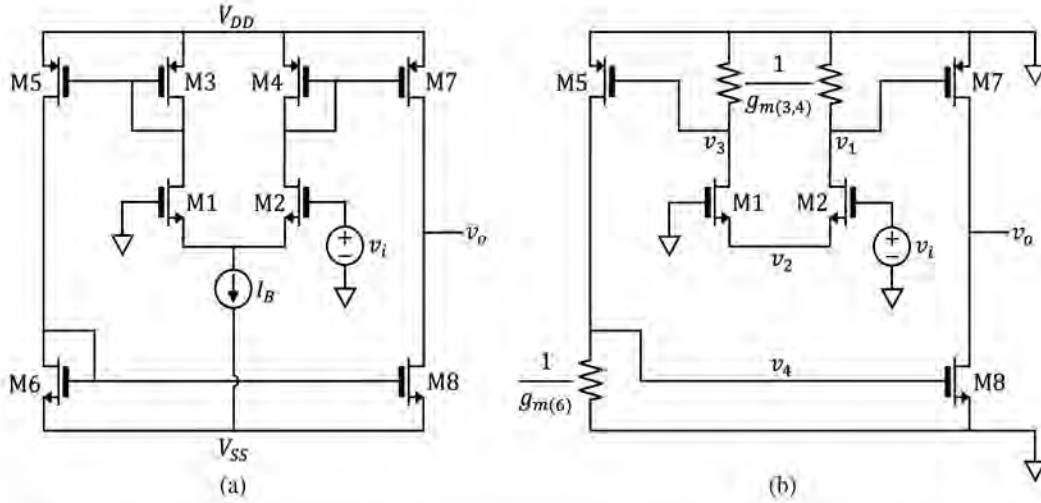


FIGURE 3.36 (a) Electrical schematic of symmetrical OTA. (b) Small-signal equivalent circuit.

each) because their dimensions are identical. Having identical bias currents, and identical overdrive voltages, their transconductances must also be identical (once more, refer to Fig. 3.24). We denote this transconductance with $g_{m(1,2)}$.

Next, consider M3 and M4. Each of these operates with a bias current $I_B/2$ imposed by M1 or M2. Also being dimensionally identical, their I_{DS} , hence overdrive voltages, are also identical. This implies identical transconductances as well. We denote this transconductance with $g_{m(3,4)}$.

The transconductances of M5, M6, M7, and M8 are also interrelated. To see why, first, notice that M5 and M7 operate with identical overdrive voltages, hence identical transconductance efficiencies because (a) gates of M3 and M5 are common, (b) gates of M4 and M7 are common, (c) all four share the same source, (d) M3 and M4 have identical gate bias voltages, and (e) the threshold voltages of all four are identical due to the common $V_{SB} = 0$ and identical L_{DR} . Therefore,

$$\frac{g_{m5}}{I_{D5}} = \frac{g_{m7}}{I_{D7}}$$

Next, consider M6 and M8. These two have a common V_{GS} and identical threshold voltages and therefore identical transconductance efficiencies. But also notice that M6 conducts the bias current I_{D5} of M5, and M8 conducts the bias current I_{D7} as M7. Hence,

$$\frac{g_{m6}}{I_{D5}} = \frac{g_{m8}}{I_{D7}}$$

Comparing this equation with the previous one, we conclude

$$\frac{g_{m5}}{g_{m6}} = \frac{g_{m7}}{g_{m8}}$$

We now move to small-signal analysis. The equivalent circuit appropriate for this purpose is shown in Fig. 3.36(b). Notice the absence of the ideal dc current source I_B , and grounded power-supply rails V_{DD} and V_{SS} . Also note that the diode-connected devices are replaced by their equivalent resistances $1/g_{m(3,4)}$ and $1/g_{m6}$, as we concluded in Example 3.11.

This equivalent circuit contains two forward signal paths, both of which begin at the gate of M2. One path moves from gate to drain through M2 and gate to drain through M7. The other moves from gate to source through M2, source to drain through M1, gate to drain through M5, and gate to drain through M8. If only the former path did exist, the output signal would be expressed as

$$v_o' = \underbrace{\frac{v_o}{v_1} \frac{v_1}{v_i}}_{A_v'} v_i.$$

Similarly, the second path would yield

$$v_o'' = \underbrace{\frac{v_o}{v_4} \frac{v_4}{v_3} \frac{v_3}{v_2} \frac{v_2}{v_i}}_{A_v''} v_i.$$

Since small-signal equivalent circuits are linear by definition, we can apply superposition to write $v_o = v_o' + v_o''$, hence

$$A_v \triangleq \frac{v_o}{v_i} = A_v' + A_v''.$$

We first determine by inspection each of the two factors of A_v' . We assume any r_{ds} to be much larger than any $1/g_m$:

$$A_v' = \underbrace{-(r_{ds7} \parallel r_{ds8}) g_{m7}}_{v_o/v_1} \times \underbrace{\frac{-g_{m(1,2)} \frac{1}{g_{m(3,4)}}}{g_{m(1,2)} + g_{mb(1,2)} r_{is1} + r_{is2}}}_{v_1/v_i},$$

where

$$r_{is1} = \frac{r_{ds1} + \frac{1}{g_{m(3,4)}}}{r_{ds1}(g_{m(1,2)} + g_{mb(1,2)})} \cong \frac{1}{g_{m(1,2)} + g_{mb(1,2)}},$$

and

$$r_{is2} = \frac{r_{ds2} + \frac{1}{g_{m(3,4)}}}{r_{ds2}(g_{m(1,2)} + g_{mb(1,2)})} \cong \frac{1}{g_{m(1,2)} + g_{mb(1,2)}}.$$

Hence,

$$A'_v = (r_{ds7} \parallel r_{ds8}) \frac{g_{m7}g_{m(1,2)}}{2g_{m(3,4)}}.$$

We now repeat the analysis for A''_v :

$$A''_v = \underbrace{-(r_{ds7} \parallel r_{ds8})g_{m8}}_{v_4/v_4} \times \underbrace{\frac{-1}{g_{m6}r_{ds5}}}_{v_4/v_3} \times \underbrace{\frac{1}{g_{m(3,4)}r_{is1}}}_{v_3/v_2} \times \underbrace{\frac{g_{m(1,2)}}{g_{m(1,2)} + g_{mb(1,2)}} \frac{r_{is1}}{r_{is1} + r_{is2}}}_{v_2/v_1},$$

where

$$r_{is5} = \frac{r_{ds5} + \frac{1}{g_{m6}}}{r_{ds5}g_{m5}} \cong \frac{1}{g_{m5}}.$$

Hence

$$A''_v = (r_{ds7} \parallel r_{ds8}) \frac{g_{m8}g_{m(1,2)}g_{m5}}{2g_{m(3,4)}g_{m6}}.$$

Considering $g_{m5}/g_{m6} = g_{m7}/g_{m8}$ as derived earlier in the bias analysis, this equation can be rewritten as

$$A''_v = (r_{ds7} \parallel r_{ds8}) \frac{g_{m7}g_{m(1,2)}}{2g_{m(3,4)}}.$$

Finally, summing up A'_v and A''_v , we obtain

$$A_v = (r_{ds7} \parallel r_{ds8}) \frac{g_{m7}g_{m(1,2)}}{g_{m(3,4)}}. \quad \blacktriangle \quad (3.70)$$

Example 3.15 Shown in Fig. 3.37(a) is a small-signal equivalent circuit with a negative feedback loop. Assuming any r_{ds} to be much larger than any $1/g_m$, determine the voltage gain v_o/v_i .

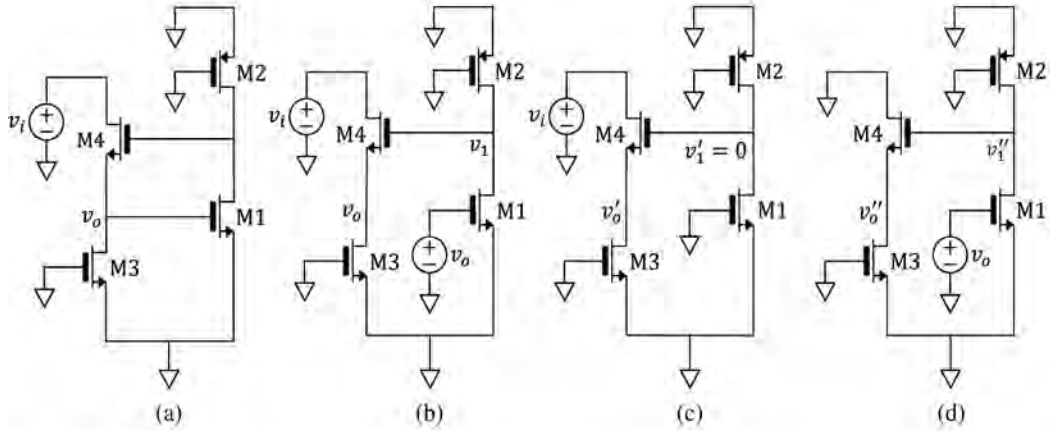


FIGURE 3.37 (a) A small-signal equivalent circuit with feedback. (b) Cutting the feedback loop at the gate of M1. (c) Equivalent circuit for determining v'_o and v'_1 when only v_i is present. (d) Equivalent circuit for determining v''_o and v''_1 when only v_o is present.

Work As always, we begin by tracing the signal path. The signal, applied to the drain terminal of M4 as v_i , flows to the source of the same device as v_o but it does not stop there. It is fed back to the gate of M4 through the gate-to-drain path of M1, and then, ends up again at the output through the gate-to-source path of M4. In order to be able to analyze the circuit by inspection, we must break this feedback loop and thus turn the circuit into an equivalent circuit containing forward paths only. Breaking at any node along the feedback path will do but, for the sake of analytical simplicity, it is most convenient to break the loop where it connects the output, that is, the gate of M1. Shown in Fig. 3.37(b) is the equivalent circuit obtained after breaking the loop. Notice the addition of a voltage source v_o to the gate of M1 for driving the broken feedback path with its original voltage. Generally, we also need to attach a resistance to the broken node of the forward path for representing the original loading of the feedback loop on this node. No such resistance is added in Fig. 3.37(b) because the load resistance seen into the gate of M1 in the original circuit of Fig. 3.37(a) is infinite.

The topology depicted in Fig. 3.37(b) has two voltage sources, v_i and v_o , but no feedback path. We now apply superposition to two forward paths, one from voltage-source v_i to output, and another from voltage-source v_o to output. First, we replace the source v_o with a short but retain the source v_i as shown in Fig. 3.37(c), and determine the corresponding output signal v'_o as a function of v_i . Inspection yields

$$v'_o = \frac{r_{ds3}}{r_{ds4} + r_{ds3}(g_{m4} + g_{mb4})r_{ds4}} v_i \approx \frac{1}{(g_{m4} + g_{mb4})r_{ds4}} v_i.$$

Next, we replace the source v_i with ground, and revive the source v_o as shown in Fig. 3.37(d), and determine the corresponding output-node signal v''_o as a function of v_o . Inspection yields

$$v''_o = -g_{m1}(r_{ds1} \parallel r_{ds2}) \times \frac{g_{m4}}{g_{m4} + g_{mb4}} \frac{r_{ds3}}{r_{ds3} + \frac{1}{g_{m4} + g_{mb4}}} v_o \approx -g_{m1}(r_{ds1} \parallel r_{ds2}) \times \frac{g_{m4}}{g_{m4} + g_{mb4}} v_o.$$

Next, we equate the sum of v'_o and v''_o to v_o , and solve for v_o to obtain

$$v_o = \frac{1}{1 + g_{m1}(r_{ds1} \parallel r_{ds2}) \frac{g_{m4}}{g_{m4} + g_{m4}}} v_i \approx \frac{1}{g_{m1}(r_{ds1} \parallel r_{ds2}) g_{m4} r_{ds4}} v_i \quad \blacktriangle$$

3.7.3 MOSFET Capacitances and High-Frequency Small-Signal Model

MOS晶体管容抗和高频小信号模型

The dielectric property of the gate-oxide layer and pn junctions gives rise to parasitic capacitances in a MOSFET structure. Capacitances confined to the channel area are called *intrinsic capacitances*. Those outside the channel area are named *extrinsic capacitances*. In this section, we model both types of parasitic capacitances, and construct a MOSFET high-frequency small-signal model by adding them to the dc small-signal equivalent circuit developed in Subsection 3.7.1.

Intrinsic Capacitances

Intrinsic capacitances are confined to the shaded channel/gate region shown on a cross-sectional view in Fig. 3.38(a). This region generally supports a total charge Q_C (C) inside the channel, and a total charge Q_G (C) inside the gate. Since electric field is assumed to vanish at the boundaries of this region, Gauss's law implies $Q_G = -Q_C$. Note that in general Q_C and therefore Q_G are functions of all four terminal voltages v_G , v_D , v_S , and v_B with respect to circuit ground. Now consider a dc steady state in which none of these terminal voltages changes. In this state, channel receives a dc current I_D across the drain boundary, and sends out the same I_D across the source boundary. Since no current flows across the bulk or gate-oxide boundaries either, no net charge is added to or extracted from the channel. This is why Q_C remains time invariant in a dc steady state. Q_C also remains time invariant because no current flows into or out of the gate.

Next, suppose that one or more of the terminal voltages are changed. In response, Q_C and therefore Q_G should also change. Consider the change in Q_C first. If it is in

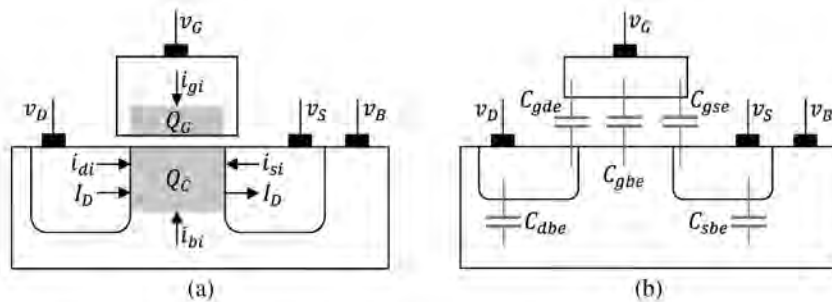


FIGURE 3.38 (a) MOSFET intrinsic capacitances are confined to the shaded channel/gate region. i_{gi} , i_{si} , i_{di} , and i_{bi} represent intrinsic charging currents flowing at the boundaries of this region. (b) Extrinsic capacitances C_{gde} , C_{gbe} , C_{gse} , C_{dbe} , and C_{sbe} .

the positive direction, a current should temporarily flow into the channel in order to bring in the additional positive charge. If the change is in the negative direction, the temporal current will flow out of the channel in order to remove the excess positive charge. In both cases, this current can flow only through the source, drain, or bulk boundaries of the channel but not through the gate-oxide boundary since the insulating property of the latter prevents charge transport. These three temporal current components flowing under ac or transient conditions are represented with i_{di} , i_{bi} , and i_{si} in Fig. 3.38(a). The channel-charge components altered by these currents are denoted by Q_D , Q_B , and Q_S , respectively. Therefore, $Q_C = Q_D + Q_B + Q_S$. Now, consider the variation in Q_C . Obviously, only a current i_{gi} temporarily flowing into or out of the gate can alter Q_C because no net charge exchange can occur through the gate-oxide boundary.

We refer to i_{gi} , i_{di} , i_{bi} , and i_{si} as *intrinsic charging current components*. For the reference directions identified in Fig. 3.38(a), they are related to the temporal derivatives of the associated charge components by the following four equations:

$$i_{gi} = \frac{dQ_C}{dt} = \underbrace{\frac{\partial Q_C}{\partial v_G} \frac{dv_G}{dt}}_{C_{gg}} + \underbrace{\frac{\partial Q_C}{\partial v_D} \frac{dv_D}{dt}}_{C_{gd}} + \underbrace{\frac{\partial Q_C}{\partial v_B} \frac{dv_B}{dt}}_{C_{gb}} + \underbrace{\frac{\partial Q_C}{\partial v_S} \frac{dv_S}{dt}}_{C_{gs}}, \quad (3.71)$$

$$i_{di} = \frac{dQ_D}{dt} = \underbrace{\frac{\partial Q_D}{\partial v_G} \frac{dv_G}{dt}}_{C_{dg}} + \underbrace{\frac{\partial Q_D}{\partial v_D} \frac{dv_D}{dt}}_{C_{dd}} + \underbrace{\frac{\partial Q_D}{\partial v_B} \frac{dv_B}{dt}}_{C_{db}} + \underbrace{\frac{\partial Q_D}{\partial v_S} \frac{dv_S}{dt}}_{C_{ds}}, \quad (3.72)$$

$$i_{bi} = \frac{dQ_B}{dt} = \underbrace{\frac{\partial Q_B}{\partial v_G} \frac{dv_G}{dt}}_{C_{bg}} + \underbrace{\frac{\partial Q_B}{\partial v_D} \frac{dv_D}{dt}}_{C_{bd}} + \underbrace{\frac{\partial Q_B}{\partial v_B} \frac{dv_B}{dt}}_{C_{bb}} + \underbrace{\frac{\partial Q_B}{\partial v_S} \frac{dv_S}{dt}}_{C_{bs}}, \quad (3.73)$$

$$i_{si} = \frac{dQ_S}{dt} = \underbrace{\frac{\partial Q_S}{\partial v_G} \frac{dv_G}{dt}}_{C_{sg}} + \underbrace{\frac{\partial Q_S}{\partial v_D} \frac{dv_D}{dt}}_{C_{sd}} + \underbrace{\frac{\partial Q_S}{\partial v_B} \frac{dv_B}{dt}}_{C_{sb}} + \underbrace{\frac{\partial Q_S}{\partial v_S} \frac{dv_S}{dt}}_{C_{ss}}. \quad (3.74)$$

These four modeling equations contain sixteen capacitances. However, only nine out of these sixteen are independent. To understand why, first, consider a specific case in which all terminal voltages change with time in an identical fashion; that is,

$$\frac{dv_G}{dt} = \frac{dv_D}{dt} = \frac{dv_B}{dt} = \frac{dv_S}{dt}, \quad (3.75)$$

which implies time-invariant port voltages V_{GS} , V_{DS} , and V_{SB} , hence time-invariant terminal currents. In other words, the time-dependent charging currents i_{gi} , i_{di} , i_{bi} , and i_{si} must vanish in this case. Replacing the left-hand sides of (3.71) to (3.74) with zero, and applying (3.75) to their right-hand sides, we obtain

$$C_{ggi} = -C_{gdi} - C_{gbi} - C_{gsi}, \quad (3.76)$$

$$C_{ddi} = -C_{dgi} - C_{dbi} - C_{dsi}, \quad (3.77)$$

$$C_{bbi} = -C_{bgi} - C_{bdi} - C_{bsi}, \quad (3.78)$$

$$C_{ssi} = -C_{sgi} - C_{sdi} - C_{sbi}. \quad (3.79)$$

Another set of relations among these capacitances is obtained by applying Kirchhoff's current law to the intrinsic volume. Irrespective of how the four terminal voltages change with time, the sum of the four charging currents must be zero all the time. Now, suppose that only v_G varies with time while v_D , v_B , and v_S are time invariant. Summing up (3.71) to (3.74) after deleting the second, third, and fourth terms on the right-hand sides, and equating the sum to zero, we find

$$C_{sgi} = -C_{ggi} - C_{dgi} - C_{bgi}. \quad (3.80)$$

Assuming only v_D to be time variant, and applying Kirchhoff's law, we obtain a second equation:

$$C_{sdi} = -C_{gdi} - C_{ddi} - C_{bdi}. \quad (3.81)$$

Repeating the procedure for the case of v_B being the sole time-variant voltage, we obtain a third equation:

$$C_{sbi} = -C_{gbi} - C_{dbi} - C_{bbi}. \quad (3.82)$$

We could repeat the procedure one more time by assuming v_S to be time-variant while the remaining terminal voltages to be constant but this would not result in an independent equation, unlike (3.76) to (3.82). These seven equations render nine of the sixteen capacitances independent. As long as any nine are available, you can calculate the rest using (3.76) to (3.82).

Equations (3.71) to (3.74) represent the intrinsic capacitive behavior in terms of terminal voltages defined with respect to circuit ground. In order to comply with our port convention set earlier in Fig. 3.2, we now convert the voltage variables of these equations into v_{GS} , v_{DS} , and v_{SB} . Actually, we need only three of these equations for constructing an equivalent circuit because Kirchhoff's current law constrains them. We select (3.71) to

(3.73) for conversion. Substituting $v_G = v_{GS} + v_S$, $v_D = v_{DS} + v_S$, $v_B = -v_{SB} + v_S$ into these three equations, and considering (3.76) to (3.78), we arrive at

$$i_{gi} = C_{ggi} \frac{dv_{GS}}{dt} + C_{gdi} \frac{dv_{DS}}{dt} - C_{gbi} \frac{dv_{SB}}{dt}, \quad (3.83)$$

$$i_{di} = C_{dgi} \frac{dv_{GS}}{dt} + C_{ddi} \frac{dv_{DS}}{dt} - C_{dbi} \frac{dv_{SB}}{dt}, \quad (3.84)$$

$$i_{bi} = C_{bgi} \frac{dv_{GS}}{dt} + C_{bdi} \frac{dv_{DS}}{dt} - C_{bbi} \frac{dv_{SB}}{dt}. \quad (3.85)$$

Finally, we rewrite these three expressions in s domain for later use in the high-frequency equivalent circuit of a MOSFET:

$$i_{gi} = sC_{ggi}v_{gs} + sC_{gdi}v_{ds} - sC_{gbi}v_{sb}, \quad (3.86)$$

$$i_{di} = sC_{dgi}v_{gs} + sC_{ddi}v_{ds} - sC_{dbi}v_{sb}, \quad (3.87)$$

$$i_{bi} = sC_{bgi}v_{gs} + sC_{bdi}v_{ds} - sC_{bbi}v_{sb}. \quad (3.88)$$

Intrinsic capacitances are bias dependent. In saturation, they vary mainly with the overdrive voltage. They are also proportional to gate area WL . Their areal density defined by

$$\overline{C_{xyi}} \triangleq \frac{C_{xyi}}{WL} \quad (3.89)$$

is, therefore, independent of device size but dependent on the overdrive voltage. We can characterize the areal densities in a device-independent manner by running a dc analysis by sweeping V_{GS} , and use the outcome for estimating the actual capacitances for any device of the same model deck. Shown in Figs. 3.39 and 3.40 are such areal-density plots (in units of $\text{fF}/\mu\text{m}^2$) of the capacitances appearing in (3.86) to (3.88). These plots belong to the NMOS and PMOS model decks of Table 3.1, and have been extracted with Spice input files `nmos2.sp` and `pmos2.sp` (see Subsections 3.3.3 and 3.4.4) after changing the `.save` lines into

```
.save @m[cgg] @m[cgd] @m[cgs] @m[cdg] @m[cdd] @m[cds]
+ @m[cbg] @m[cdb] @m[cbs] @m[vth] v(1,3)
```

Except for $\overline{C_{gbi}}$, $\overline{C_{dbi}}$, and $\overline{C_{bbi}}$, areal densities in units of $\text{fF}/\mu\text{m}^2$ are plotted with the command

```
plot @m[cxy]/4.6e-15 vs v(1,2)-@m[vth]
```

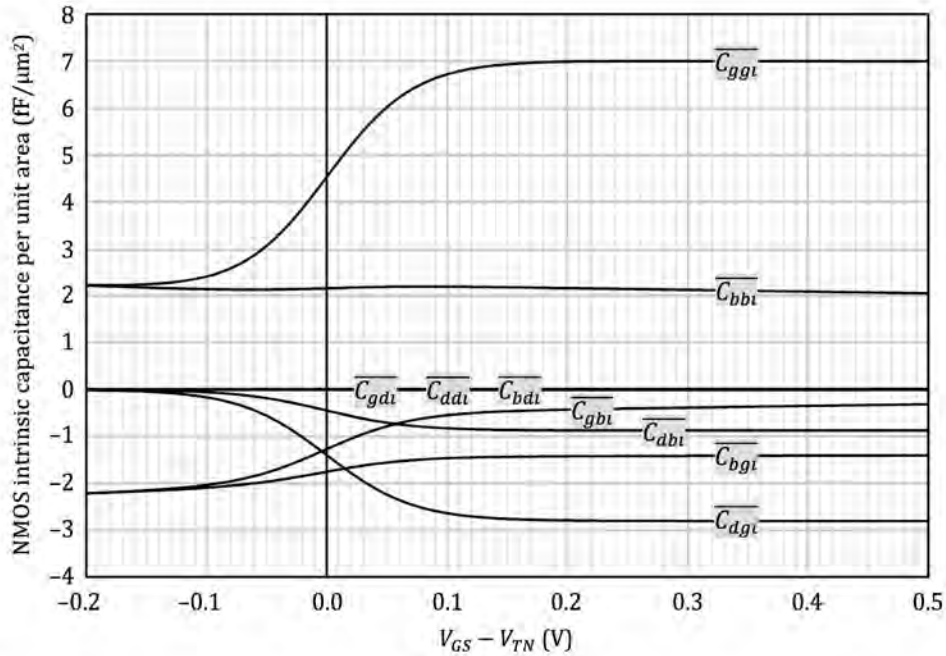


FIGURE 3.39 Variation of the density (per unit channel area) of intrinsic capacitances with overdrive voltage in an NMOS in saturation.

for NMOS whose effective channel area equals $4.60 \mu\text{m}^2$, and with the command

```
plot @m[cxy]/4.44e-15 vs v(1,3)-@m[vth]
```

for PMOS whose effective channel area equals $4.44 \mu\text{m}^2$. Since C_{gbi} , C_{dbi} , and C_{bbi} are not directly accessible in the particular version of Ngspice used in these simulations, their areal densities have been plotted for NMOS with

```
plot (-@m[cgg]-@m[cgd]-@m[cgs])/4.6e-15 vs v(1,3)-@m[vth]
plot (-@m[cdd]-@m[cdg]-@m[cds])/4.6e-15 vs v(1,3)-@m[vth]
plot (-@m[cbg]-@m[cbd]-@m[cbs])/4.6e-15 vs v(1,3)-@m[vth]
```

as implied by (3.76) to (3.78), respectively. Same entries have been used for PMOS counterparts after changing the factor of normalization from 4.6×10^{-15} to 4.44×10^{-15} .

Notice the following features of the areal density plots observed in Figs. 3.39 and 3.40 for a device operating in saturation:

- Some of the capacitances have negative values. This is a result of the reference directions implicitly assumed in defining intrinsic capacitances in (3.71) to (3.74). As an example, consider C_{gbi} . It is defined as $C_{gbi} \doteq \partial Q_G / \partial v_B$ in (3.71). Physically, an increasing v_B causes the gate charge Q_G to decrease. Therefore, the partial derivative is negative.
- $\overline{C_{gdi}}$, $\overline{C_{ddi}}$, and $\overline{C_{bdi}}$ are negligibly small.

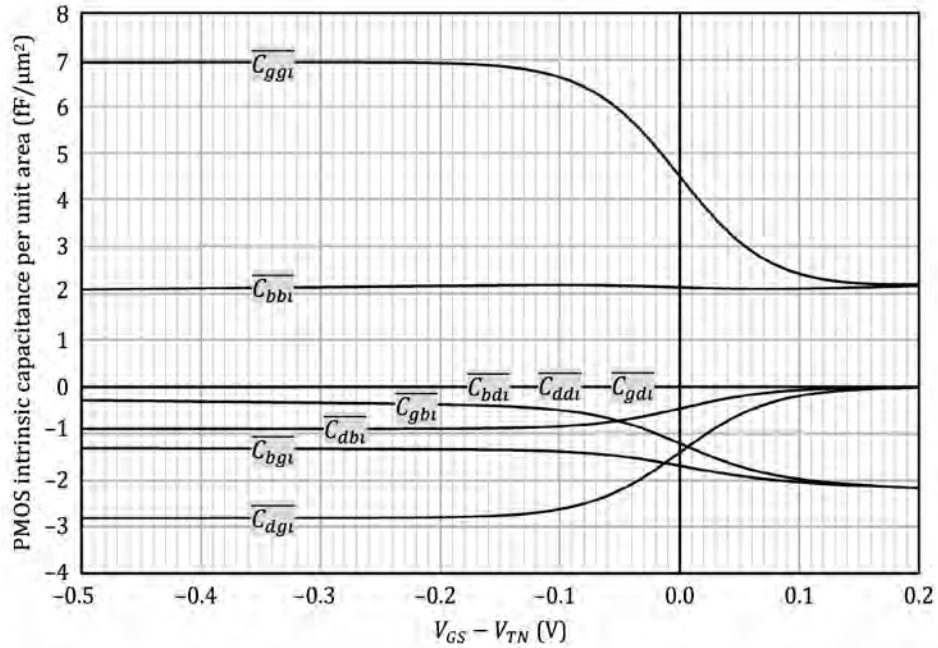


FIGURE 3.40 Variation of the intrinsic capacitances per unit gate area with overdrive voltage in a PMOS in saturation.

- $\overline{C_{bbi}}$ is virtually constant regardless of the strength of inversion.
- $\overline{C_{ggi}}$ is virtually independent of the overdrive voltage in strong inversion for approximately $V_{GS} - V_{TN} > 0.1$ V in NMOS and $V_{GS} - V_{TP} < -0.1$ V in PMOS. It decreases as the device moves into weak inversion and subthreshold.
- $\overline{C_{gbi}}$ and $\overline{C_{bgi}}$ are virtually independent of the overdrive voltage in strong inversion for approximately $V_{GS} - V_{TN} > 0.1$ V in NMOS and $V_{GS} - V_{TP} < -0.1$ V in PMOS. Their magnitude increases as the device moves into weak inversion and subthreshold.
- $\overline{C_{dbi}}$ and $\overline{C_{dgi}}$ are virtually independent of the overdrive voltage in strong inversion for approximately $V_{GS} - V_{TN} > 0.1$ V in NMOS and $V_{GS} - V_{TP} < -0.1$ V in PMOS. Their magnitude decreases as the device moves into weak inversion and subthreshold.

As a final word, note that $\overline{C_{ggi}}$ happens to be the most influential intrinsic capacitance in ac or transient behavior of a MOSFET. It is worth remembering that its areal density is reasonably well approximated by

$$\overline{C_{ggi}} \approx \frac{2}{3} C_{ox} \quad (3.90)$$

under strong-inversion conditions in saturation.

Extrinsic Capacitances

The extrinsic capacitances of a MOSFET are represented in Fig. 3.38(b) with C_{gbe} , C_{gde} , C_{gse} , C_{dbe} , and C_{sbe} . We now describe the structural and electrical variables of each of these capacitances.

C_{gbe} is due to the fringing electrical field between gate and bulk. It is independent of bias conditions but is proportional to channel length as described by

$$C_{gbe} = L \times CGBO, \quad (3.91)$$

where $CGBO$ (F/m) is a BSIM parameter. C_{gbe} is usually small enough to be ignored in design.

C_{gde} is due to the overlap between gate and drain junction. It is independent of bias conditions but is proportional to channel width as described by

$$C_{gde} = W \times CGDO, \quad (3.92)$$

where $CGDO$ (F/m) is a BSIM parameter.

C_{gse} is due to the overlap between gate and source junction. It is also independent of bias conditions and proportional to channel width, and is described by

$$C_{gse} = W \times CGSO, \quad (3.93)$$

where $CGSO$ (F/m) is a BSIM parameter.

C_{dbe} is the junction capacitance associated with drain-bulk junction. It has three components, all of which depend on drain-bulk bias V_{DB} as follows:

$$C_{dbe} = \frac{A_D \times CJ}{\left(1 + \frac{|V_{DB}|}{PB}\right)^{MJ}} + \frac{(P_D - W) \times CJSW}{\left(1 + \frac{|V_{DB}|}{PBSW}\right)^{MJSW}} + \frac{W \times CJSWG}{\left(1 + \frac{|V_{DB}|}{PBSWG}\right)^{MJSWG}}, \quad (3.94)$$

where A_D (m²) is the area of the drain-bulk junction, and P_D (m) is the total perimeter of the same junction. In a Spice input file, these two parameters are specified as instance parameters in the element line of the MOSFET they belong to. Among the BSIM3 parameters appearing in (3.94), MJ, MJSW, and MJSWG are dimensionless. The rest are included in the model deck with the following units: CJ in F/m²; CJSW and CJSWG in F/m; PB, PBSW, and PBSWG in V.

C_{sbe} is the junction capacitance associated with the source-bulk junction. It has three components, all of which depend on the source-bulk bias V_{SB} as follows:

$$C_{sbe} = \frac{A_S \times CJ}{\left(1 + \frac{|V_{SB}|}{PB}\right)^{MJ}} + \frac{(P_S - W) \times CJSW}{\left(1 + \frac{|V_{SB}|}{PBSW}\right)^{MJSW}} + \frac{W \times CJSWG}{\left(1 + \frac{|V_{SB}|}{PBSWG}\right)^{MJSWG}}, \quad (3.95)$$

where A_j (m^2) is junction area and P_j (m) is total junction perimeter. In a Spice input file, these two parameters are also specified as instance parameters in MOSFET element line.

In (3.94) and (3.95), the first term on the right-hand side represents the capacitance of the bottom surface of the junction, whereas the second and third terms are the junction sidewall capacitances outside the channel and inside the channel, respectively.

Having modeled each of the five extrinsic capacitances, we now express the *extrinsic charging currents* i_{ge} , i_{de} , and i_{be} flowing into the device at the gate, drain, and bulk terminals due to these capacitances:

$$\begin{aligned} i_{ge} &= C_{gde} \frac{dv_{GD}}{dt} + C_{gse} \frac{dv_{GS}}{dt} + C_{gbe} \frac{dv_{GB}}{dt}, \\ i_{de} &= -C_{gde} \frac{dv_{GD}}{dt} + C_{dbe} \frac{dv_{DB}}{dt}, \\ i_{be} &= -C_{gbe} \frac{dv_{GB}}{dt} - C_{sbe} \frac{dv_{SB}}{dt} - C_{dbe} \frac{dv_{DB}}{dt}. \end{aligned}$$

Substituting $v_{GS} - v_{DS}$ for v_{GD} , $v_{GS} + v_{SB}$ for v_{GB} , and $v_{DS} + v_{SB}$ for v_{DB} , we rewrite these three expressions in terms of our conventional port voltages v_{GS} , v_{DS} , and v_{SB} :

$$i_{ge} = C_{ggs} \frac{dv_{GS}}{dt} - C_{gde} \frac{dv_{DS}}{dt} + C_{gbe} \frac{dv_{SB}}{dt}, \quad (3.96)$$

$$i_{de} = -C_{gde} \frac{dv_{GS}}{dt} + C_{dde} \frac{dv_{DS}}{dt} + C_{dbe} \frac{dv_{SB}}{dt}, \quad (3.97)$$

$$i_{be} = -C_{gbe} \frac{dv_{GS}}{dt} - C_{dbe} \frac{dv_{DS}}{dt} - C_{bbe} \frac{dv_{SB}}{dt}, \quad (3.98)$$

where

$$C_{ggs} \doteq C_{gse} + C_{gde} + C_{gbe}, \quad (3.99)$$

$$C_{dde} \doteq C_{gde} + C_{dbe}, \quad (3.100)$$

$$C_{bbe} \doteq C_{gbe} + C_{dbe} + C_{sbe}. \quad (3.101)$$

Finally, we convert (3.96) to (3.98) into s -domain expressions:

$$i_{ge} = sC_{ggs}v_{gs} - sC_{gde}v_{ds} + sC_{gbe}v_{sb}, \quad (3.102)$$

$$i_{de} = -sC_{gde}v_{gs} + sC_{dde}v_{ds} + sC_{dbe}v_{sb}, \quad (3.103)$$

$$i_{be} = -sC_{gbe}v_{gs} - sC_{dbe}v_{ds} - sC_{bbe}v_{sb}, \quad (3.104)$$

High-Frequency Equivalent Circuit

The total capacitive charging current flowing into the device at a given terminal is the sum of intrinsic and extrinsic components. Using (3.86) to (3.88) and (3.102) to (3.104) for these components, we can express the total charging currents as follows

$$i_g = i_{gi} + i_{ge} = s \underbrace{(C_{gge} + C_{ggi})}_{C_{gg}} v_{gs} - s \underbrace{(C_{gde} - C_{gdi})}_{C_{gd}} v_{ds} + s \underbrace{(C_{gbe} - C_{gbi})}_{C_{gb}} v_{sb},$$

$$i_d = i_{di} + i_{de} = -s \underbrace{(C_{gde} - C_{gdi})}_{C_{dg}} v_{gs} + s \underbrace{(C_{dde} + C_{ddi})}_{C_{dd}} v_{ds} + s \underbrace{(C_{dbe} - C_{dbi})}_{C_{db}} v_{sb},$$

$$i_b = i_{bi} + i_{be} = -s \underbrace{(C_{gbe} - C_{gbi})}_{C_{bg}} v_{gs} - s \underbrace{(C_{dbe} - C_{dbi})}_{C_{bd}} v_{ds} - s \underbrace{(C_{bbe} - C_{bbi})}_{C_{bb}} v_{sb},$$

where we define

$$C_{gg} \triangleq C_{gge} + C_{ggi}, \quad (3.105)$$

$$C_{gd} \triangleq C_{gde} - C_{gdi}, \quad (3.106)$$

$$C_{gb} \triangleq C_{gbe} - C_{gbi}, \quad (3.107)$$

$$C_{dd} \triangleq C_{dde} + C_{ddi}, \quad (3.108)$$

$$C_{dg} \triangleq C_{gde} - C_{gdi}, \quad (3.109)$$

$$C_{db} \triangleq C_{dbe} - C_{dbi}, \quad (3.110)$$

$$C_{bb} \triangleq C_{bbe} + C_{bbi}, \quad (3.111)$$

$$C_{bg} \triangleq C_{gbe} - C_{gbi}, \quad (3.112)$$

$$C_{bd} \triangleq C_{dbe} - C_{dbi}. \quad (3.113)$$

Finally, supplementing the dc equivalent circuit of Fig. 3.26 with these capacitive components leads us to the high-frequency small-signal equivalent circuit shown in Fig. 3.41.

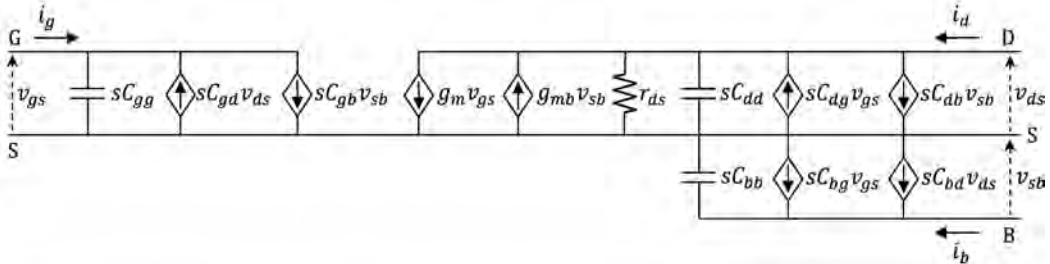


FIGURE 3.41 High-frequency small-signal equivalent circuit of a MOSFET.

Note that among nine capacitive components only C_{gg} , C_{dd} , and C_{bb} are represented by a capacitance symbol. Each of the remaining six capacitances relates the capacitive current of one port to the variation in the voltage of another, and therefore, does not fit into the proper definition of a capacitance. This is why such components are referred to as *transcapacitance*. They function as a voltage-controlled current source, and this is how they are represented in the equivalent circuit.

Example 3.16 This is our fourth visit to the common-source amplifier stage of Fig. 3.22 (after Examples 3.5, 3.10, and 3.12). This time we load its output with a capacitance $C_L = 150$ fF as shown in Fig. 3.42(a), and analyze its frequency response in terms of the voltage gain $A_v \equiv v_o/v_i$ and input admittance $Y_i \equiv i_i/v_i$ around an output bias level $V_O = 0.9$ V. For both devices, we assume a rectangular drain-bulk junction geometry of side lengths W_{DR} and 0.55 μm . As usual, we verify the results with a Spice ac analysis.

Work For the purpose of small-signal analysis, the source and bulk terminals of M1 and the source, bulk, and gate terminals of M2 are at signal ground. Using the MOSFET equivalent circuit of Fig. 3.41 for both devices after grounding these terminals, we represent the amplifier stage with the equivalent circuit shown in Fig. 3.42(b). Notice that only r_{ds2} and C_{dd2} are included from the equivalent circuit of M2. This is a consequence of both v_{sb2} and v_{gs2} being zero. Therefore, M2 acts merely as a parallel equivalent of resistance and a capacitance. Straightforward circuit analysis yields

$$A_v \triangleq \frac{v_o}{v_i} = -g_{m1} r \frac{1 - s \frac{C_{dg1}}{g_{m1}}}{1 + s C_L r}, \quad (3.114)$$

for the transfer function of voltage gain, and

$$Y_i \triangleq \frac{i_i}{v_i} = s \left(C_{gg1} + g_{m1} r \frac{1 - s \frac{C_{dg1}}{g_{m1}}}{1 + s C_L r} C_{gd1} \right) = s \left(C_{gg1} - A_v C_{gd1} \right), \quad (3.115)$$

for input admittance. Note the definitions $C'_L \triangleq C_L + C_{dd1} + C_{dd2}$ and $r \triangleq r_{ds1} \parallel r_{ds2}$.

For proper interpretation of (3.114) and (3.115), we need to extract the numerical values of the device parameters appearing in these expressions. This calls for a Spice operation-point analysis. But, first, we perform a dc-sweep for accurately determining the input bias voltage V_I for which the specified $V_O = 0.9$ V is realized. Accuracy is essential because even a slight deviation from the necessary input bias will shift V_O far away from the specified 0.9 V due to the large gain of the stage. We use the .dc line of the following input file, and determine the necessary input bias voltage as $V_I = 567$ mV.⁶

⁶First, plot V_O versus V_I by typing `plot v(2)`. Then, point $V_O = 0.9$ V on the transfer characteristic, and left click once. Spice puts out the corresponding V_I as `x0 = 0.567619`.

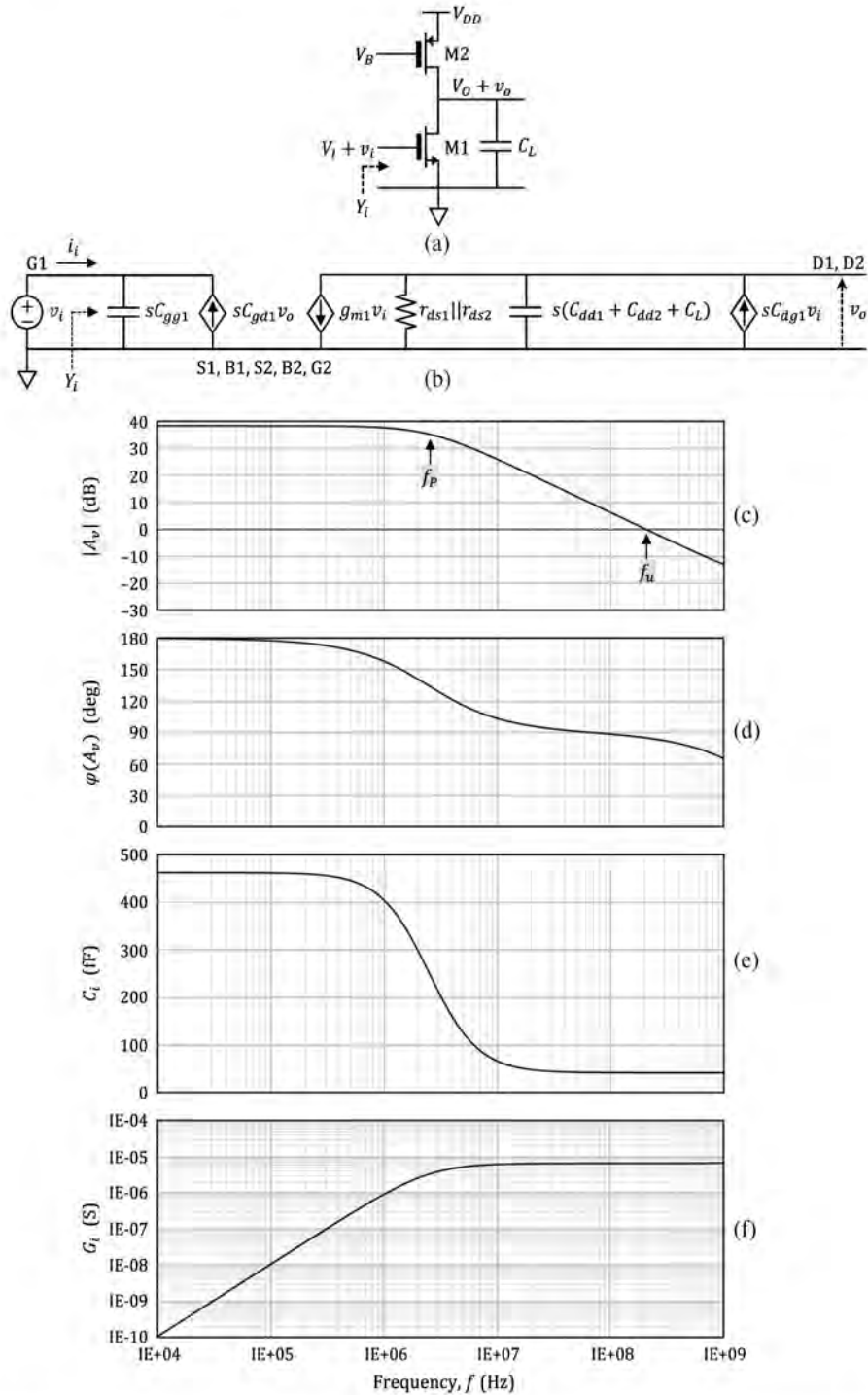


FIGURE 3.42 (a) Common-source amplifier stage loaded by a capacitance C_L . (b) High-frequency small-signal equivalent circuit for analyzing the voltage gain A_v and input admittance Y_i . (c) Magnitude response of the gain. (d) Phase response of the gain. (e) Input capacitance. (f) Input conductance.

```

filename: ex3-16.sp
.incl bsim.sp
m1 2 1 0 0 cmosn w=6u l=0.8u ad=3.3p pd=13.1u
+ nrd=3.33e-2 nrs=3.33e-2
m2 2 3 4 4 cmosp w=24u l=0.8u ad=13.2p pd=49.1u
+ nrd=8.33e-3 nrs=8.33e-3
vb 3 0 1.25
vdd 4 0 1.8
cl 2 0 0.15p
vi 1 0 dc 567m ac 1
.dc vi 0.5 0.6 1m
*.op
*.tf v(2) vi
*.ac dec 100 10k 1g
*.noise v(2) vi dec 100 10k 1g ; used in Example 3.17
.end

```

Before proceeding to the operation-point analysis, notice the declaration of the area (ad) and perimeter (pd) of the drain-bulk junctions of both MOSFETs as instance parameters. As pointed out in Subsection 3.2.1, these are necessary for the simulator to properly calculate the extrinsic drain-bulk capacitances. No area or perimeter is declared for source-bulk junctions because these junctions do not support any signal in the present circuit. Also notice the declaration of nrd and nrs as estimated by assuming a minimum contact-to-gate spacing $L_{CG} = 0.2 \mu\text{m}$ as per Table 3.2. This is recommended in any ac simulation where input admittance is to be evaluated because otherwise the resistances placed by the simulator in series with the driver device would be unrealistically large. This would lead to an inaccurate evaluation of the input admittance at high frequencies. Although declaring these two parameters only for the driver device will suffice, it is prudent to follow the practice for all MOSFETs as a precaution against possible oversight in identifying the devices that are influential in this respect.

Now, we run an operation-point simulation, and read the following parameter values for M1 with the command show: $g_{m1} = 244 \mu\text{A/V}$, $r_{ds1} = 693 \text{ k}\Omega$, $C_{ggi1} = 31.79 \text{ fF}$, $C_{gdi1} = 69 \text{ aF}$, $C_{ddi1} = -38 \text{ aF}$, $C_{dgi1} = -12.59 \text{ fF}$, and $C_{dbe1} = 5.32 \text{ fF}$ for M1.⁷ For M2, we read $r_{ds2} = 667 \text{ k}\Omega$, $C_{ddi2} = -42 \text{ aF}$, and $C_{dbe2} = 24.3 \text{ fF}$.

Next, we calculate the extrinsic capacitance $C_{gse1} = 5.15 \text{ fF}$ from (3.93), $C_{gde1} = 5.15 \text{ fF}$ and $C_{gde2} = 18.8 \text{ fF}$ from (3.92), and $C_{gbe1} = 0.8 \text{ aF}$ from (3.91). These numbers lead to $C_{gge1} = 10.3 \text{ fF}$ from (3.99), and $C_{dde1} = 10.5 \text{ fF}$ and $C_{dde2} = 43.1 \text{ fF}$ from (3.100).

Using these parameter values in (3.105), (3.106), (3.108), and (3.109), we obtain the following values for the capacitances appearing in (3.114) and (3.115): $C_{gg1} = 42.1 \text{ fF}$, $C_{gd1} = 5.08 \text{ fF}$, $C_{dd1} = 10.5 \text{ fF}$, $C_{dd2} = 43.1 \text{ fF}$, and $C_{dg1} = 17.7 \text{ fF}$. Calculation also yields $C'_L = 204 \text{ fF}$ and $r = 340 \text{ k}\Omega$.

Now that all the necessary numerical data are available to us, we can start interpreting (3.114) and (3.115). According to (3.114), voltage gain A_v has a dc value

$$A_{v0} = -g_{m1}r = -83 \text{ V/V}$$

⁷ Drain-bulk extrinsic capacitance is represented with parameter capbd in Ngspice. Source-bulk counterpart is capbs.

corresponding to 38.4 dB, which can be verified by running the transfer-function analysis included in the input file, and reading `transfer_function` after typing `print all`. A_v also has a right-half-plane zero at the frequency

$$f_z = \frac{g_{m1}}{2\pi C_{dg1}} = 2.19 \text{ GHz},$$

and a pole at the frequency

$$f_p = \frac{1}{2\pi r C'_L} = 2.29 \text{ MHz}. \quad (3.116)$$

The magnitude and phase responses of A_v are shown in Figs. 3.42(c) and (d), respectively. These are obtained with the ac analysis included in the input file. Simply, run the simulation, set the unit of phase to degrees with

```
set units=degrees
```

and subsequently plot with

```
plot vdb(2)-vdb(1) vp(2)-vp(1)
```

Also notice a unity-gain frequency f_u around 206 MHz. Since the zero is located at a much higher frequency, the amplifier exhibits a single-pole characteristic all the way up to f_u . The phase equals 180° at low frequencies but drops to 135° at $f = f_p$ as expected from the left half-plane pole. Had there been no zero, the phase would settle at 90° at high frequencies. However, it decreases further at higher frequencies under the influence of the right half-plane zero, and reaches 75° at $f = f_u$.

The input admittance described by (3.115) has a capacitive component C_i and a conductive component G_i . Both vary with frequency as shown in Figs. 3.42(e) and (f) with simulated plots. These plots are generated by typing

```
plot imag(-i(vi)/v(1))/(2*pi*frequency)
```

for C_i , and

```
plot real(-i(vi)/v(1)) loglog
```

for G_i after running the ac analysis. A straightforward piecewise analysis of (3.115) in frequency domain yields the following asymptotic values for the two components:

$$C_i = \begin{cases} C_{gg1} + g_{m1} r C_{gd1} & \text{for } f \ll f_p \ll \sqrt{f_p f_z} \\ C_{gg1} + g_{m1} r C_{gd1} \left(\frac{f_p}{f}\right)^2 & \text{for } f_p \ll f \ll \sqrt{f_p f_z} \\ C_{gg1} - \frac{C_{dg1}}{C'_L} C_{gd1} & \text{for } f_p \ll \sqrt{f_p f_z} \ll f \end{cases} \quad (3.117)$$

$$G_i = \begin{cases} g_{m1} \frac{C_{gd1}}{C'_L} \left(\frac{f}{f_p}\right)^2 & \text{for } f \ll f_p \\ g_{m1} \frac{C_{gd1}}{C'_L} & \text{for } f_p \ll f \end{cases}$$

For $f \ll f_p$, (3.117) indicates a frequency-independent C_i comprising not only C_{gg1} (42.1 fF) but also C_{gd1} (5.08 fF) multiplied by the dc gain $g_{m1}r$ (83 V/V). This dominating contribution of C_{gd1} to input capacitance is known as *Miller effect*, which is represented in the equivalent circuit of Fig. 3.42(a) by the voltage-dependent current source $sC_{gd1}v_o$ attached to the input port. Since v_o equals $A_v v_i$, the current of this source is essentially controlled by v_i , and therefore the source itself is equivalent to an admittance $-sA_v C_{gd1}$ shunting the input port. Its low-frequency form $sg_{m1}rC_{gd1}$ based on $A_v = A_{v0} = -g_{m1}r$ is indeed capacitive. However, as f rises above f_p , and therefore the gain becomes $A_v \approx -g_{m1}/sC'_L$ as approximated from (3.114), the admittance due to Miller effect turns into $g_{m1}C_{gd1}/C'_L$, which, obviously, is conductive. ▲

3.8 MOSFET Noise Model MOS晶体管噪声模型

Two distinct types of inherent noise are dominant in MOSFET devices belonging to the technology nodes covered by BSIM3v3. The random motion of thermally agitated carriers inside the channel is the underlying mechanism of one of these two. It gives rise to thermal noise, which has a frequency-independent flat spectrum exactly like its counterpart observed in resistors. The other type arises from the random fluctuation of carrier mobility and/or the random capture and release of carriers by traps located at or near the channel-oxide interface. This type of noise is called *flicker noise*, which is also known as *1/f noise* because its power spectral density (PSD) is inversely proportional to the frequency f . The effect of each is represented by a current source placed in parallel with the drain-source port as shown in Fig. 3.43(a).

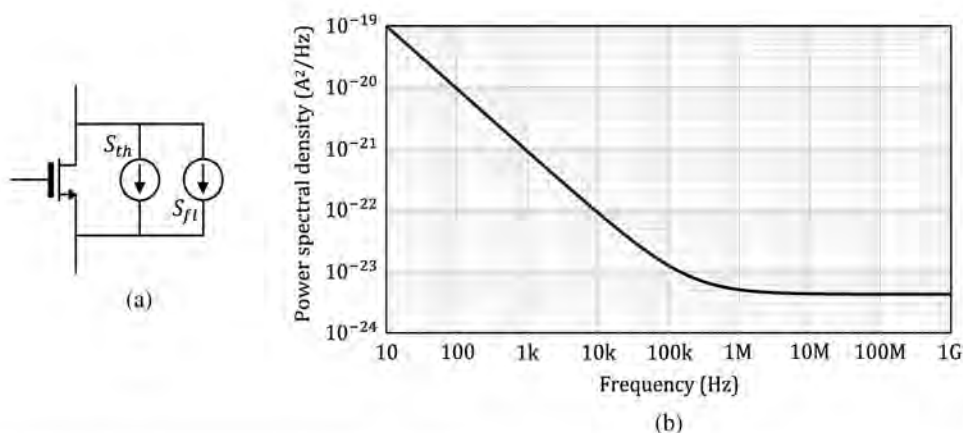


FIGURE 3.43 (a) Representing noise with independent current sources in a MOSFET, where S_{th} and S_{fl} are the power spectral densities of the thermal and flicker noise components. (b) A typical variation of S_{th} + S_{fl} with frequency.

Thermal Noise

BSIM3v3 offers two PSD models for thermal noise current. The simpler model, known as *Spice2 model* and identified by `noimod=1` or `noimod=3` in the model deck, is described by

$$S_{it} = \frac{8}{3} E_{th} (g_m + g_{mb} + g_{ds}) \quad (3.118)$$

where E_{th} is the thermal energy defined right after (2.108). This is also how we model S_{it} for manual analysis and design purposes throughout this book. Note that g_{mb} has to be retained in this expression even if the device operates with a source-bulk short but g_{ds} may be ignored in saturation. The more complex model, known as *charge-based model* and identified by `noimod=2` or `noimod=4`, is not amenable to manual analysis or design. The reader is referred to [5] for its details. In BSIM4 and BSIM-BULK the Spice2 model is abandoned. The charge-based model is retained and identified by `tnoimod=0`. A new *holistic model* is also introduced in these later BSIM versions.[6][7]

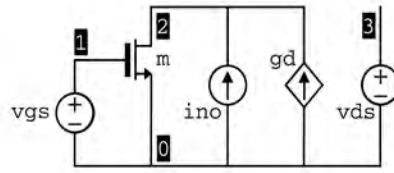
Flicker Noise

BSIM3v3 also offers two options for the PSD of flicker noise current. Model identifiers are `noimod=1` or `noimod=4` for the simple Spice2 option, and `noimod=2` or `noimod=3` for the complex charge-based option. These two models are retained in BSIM4 with identifiers `fnoimod=0` and `fnoimod=1`, respectively.[6] BSIM-BULK offers only the complex model with the same `fnoimod=1` identifier.[7] Since the complex model is what model decks usually feature, we capture its description of the PSD S_{fi} in terms of design variables with

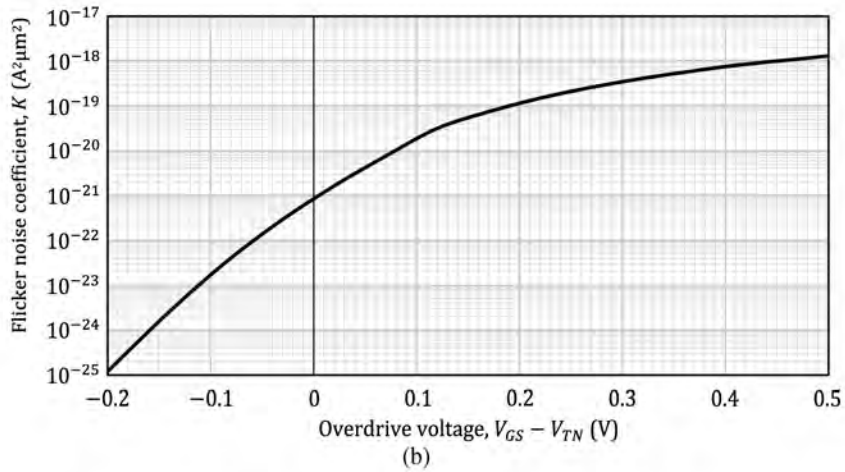
$$S_{fi} = K \frac{W}{fL^3} \quad (3.119)$$

where K is a coefficient depending only on the overdrive voltage for a given model deck. This dependence resembles that of $I_{\square}$ in strong inversion and $I_{\square}^2$ in deep subthreshold. We can characterize it with a single simulation as a function of the overdrive voltage for any given model deck. A schematic appropriate for this purpose is shown in Fig. 3.44(a) for an NMOS device. While the gate-source port is biased by an independent voltage source `vgs`, the drain-source port is regulated by a voltage-controlled dc current source `gd` placed inside a negative feedback loop. The very large transconductance gain of `gd` establishes a virtual short between the drain terminal of the MOSFET and a reference voltage source `vds`. The drain-source voltage is therefore kept constant at this `vds` while the MOSFET is automatically supplied with the bias current demanded by the overdrive voltage. This indirect biasing of the drain-source port is imperative because direct biasing with a dc voltage source would short-circuit the ac current source `ino` attached to the drain for measuring the noise current of the MOSFET. We will use this schematic for determining S_{fi} at one frequency only, and then divide the outcome by W/fL^3 to obtain K . The input file `nmos-flicker.sp` corresponding to the schematic is as follows:

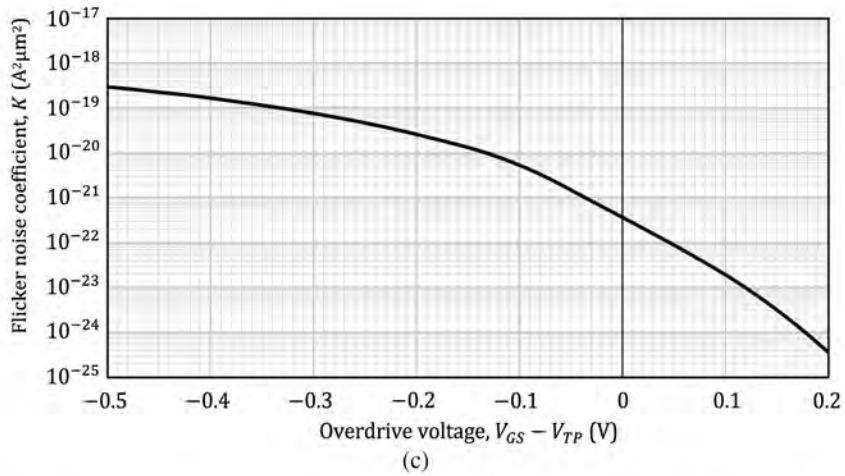
```
filename: nmos-flicker.sp
.incl bsim.sp
m 2 1 0 0 cmosn w=6u l=0.8u nrd=0 nrs=0
vgs 1 0 dc 0.6
```



(a)



(b)



(c)

FIGURE 3.44 (a) Schematic for extracting the flicker noise coefficient K for NMOS devices. (b) K as a function of overdrive voltage in NMOS devices. (c) K as a function of overdrive voltage in PMOS devices.

```

gd 0 2 3 2 1e6
vds 3 0 0.9
ino 0 2 dc 0 ac 1
.control
destroy all
foreach gate 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5
+0.55 0.6 0.65 0.7 0.75 0.8 0.85 1.0
alter vgs=$gate
noise v(2) ino dec 1 10 10
end
wrrdata nmos-flicker
+ noise1.inoise_spectrum noise2.inoise_spectrum
+ noise3.inoise_spectrum noise4.inoise_spectrum
+ noise5.inoise_spectrum noise6.inoise_spectrum
+ noise7.inoise_spectrum noise8.inoise_spectrum
+ noise9.inoise_spectrum noise10.inoise_spectrum
+ noise11.inoise_spectrum noise12.inoise_spectrum
+ noise13.inoise_spectrum noise14.inoise_spectrum
+ noise15.inoise_spectrum noise16.inoise_spectrum
+ noise17.inoise_spectrum @m[vth]
.endc
.end

```

It is worth noting the following key points of this input file:

- Noise analysis is described by the line

```
noise v(2) ino dec 1 10 10
```

where, quite generally, the first entry following `noise` specifies an *output port* with a voltage, which happens to be `v(2)` in the present case. The next entry is an *ac input source*, which may be a current source as identified with `ino` in the present case or a voltage source. Spice determines not only the noise voltage across the output port but also the equivalent noise voltage or noise current the input source would assume in order to generate the very same noise voltage across the output port in a noise-free circuit. These equivalent noise entities are known as *input-referred* noise voltage or current. Since Spice does not support noise current at the output port we define `ino` as an input current source and attach it to the drain of the MOSFET where we would like to characterize the PSD. Output port definition is irrelevant for our purposes, which is why its definition `v(2)` is quite arbitrary.

- Since we need to determine S_{η} only at a single frequency, the start and stop frequencies of the noise analysis are set to the same 10 Hz. The reason why we select $f = 10$ Hz is that it is low enough to ensure $S_{\eta} \gg S_{th}$ so that the PSD determined by simulation can be attributed almost entirely to flicker noise.
- `nrs` and `nrd` are set to zero in order to not involve the noise due to the parasitic resistances of the source and drain terminals of the MOSFET.
- Noise analysis is repeated for 17 different values of the gate-source voltage as defined inside the `foreach-end` loop. These analyses are numbered in ascending order starting with 1 for $V_{GS} = 0.1$ V and ending with 17 for $V_{GS} = 1.0$ V.

The square root $S_{\text{fl}}^{1/2}$ of the PSD of noise current flowing in ino and therefore in the MOSFET is identified by

```
noiseX.inoise_spectrum
```

where $X = [1, \dots, 17]$ is the order of the analysis in which it is determined. These values of $S_{\text{fl}}^{1/2}$ are written into the data file `nmos-flicker` as instructed with the `wrdata` line. Each such value, preceded by an entry of the frequency 10 Hz at which the simulation has been run, will appear in a data string inside this data file. The threshold voltage `@m[vth]` will appear as the last entry of the string. We process this data by (a) copying the entire string to a spreadsheet software, (b) deleting all 10-Hz entries, (c) squaring the PSD data to obtain S_{fl} , and dividing the outcome by

$$\frac{W}{fL^3} = \frac{W_{\text{DR}} + XW - 2 \times \text{WINT}}{f(L_{\text{DR}} + XL - 2 \times \text{LINT})^3} = 1.33 \frac{1}{\text{Hz} \cdot \mu\text{m}^2},$$

as stated just before the input file, and (d) pairing each PSD value with the corresponding overdrive voltage calculated from the corresponding $v_{\text{GS}} - @m[vth]$. Plotting the outcome yields K in units of $\text{A}^2\mu\text{m}^2$ as a function of $V_{\text{GS}} - V_{\text{TN}}$. The plot, presented in Fig. 3.44(b), is valid for any NMOS device of the model deck `bsim.sp`.

The reader is now invited to construct a circuit schematic and the corresponding Spice input file for extracting K for PMOS devices. For our exemplar model deck `bsim.sp`, simulation should yield the plot shown in Fig. 3.44(c).

Power Spectral Density of Total Noise Current

The schematic given in Fig. 3.44(a) and its PMOS counterpart are also good for characterizing the PSD $S_{\text{th}} + S_{\text{fl}}$ of the total noise current of a MOSFET as a function of frequency for a given bias condition. The following input file `nmos-noise.sp` is an example serving this purpose for an NMOS device:

```
filename: nmos-noise.sp
.incl bsim.sp
m 2 1 0 0 cmosn w=6u l=0.8u nrd=0 nrs=0
vgs 1 0 dc 0.6
gd 0 2 3 2 1e6
vds 3 0 0.9
ino 0 2 dc 0 ac 1
.control
destroy all
noise v(2) ino dec 100 10 1g
setplot noise1
plot inoise_spectrum^2 loglog
.endc
.end
```

The outcome of the simulation run with this file is plotted in Fig. 3.43(b). Notice the dominance of the flicker noise at low frequencies and of the thermal noise at high frequencies.

Example 3.17 Determine the root mean square (rms) output noise voltage of the common-source amplifier stage of Example 3.16 for $f > 10$ Hz at room temperature. Verify the result with a Spice simulation.

Work Shown in Fig. 3.45(a) is the circuit schematic of the amplifier, where $i_{n1}(t)$ and $i_{n2}(t)$ represent the noise current sources of the two devices. We first need to derive the transfer function from each of these sources to the output noise voltage v_n . To that end, we construct the high-frequency equivalent circuit corresponding to this circuit schematic. We do this by referring back to Fig. 3.42(b) of Example 3.16, and (a) replacing the input signal source v_i with a short circuit (assuming zero-source resistance), and (b) representing the parallel equivalent of C_{dd1} , C_{dd2} , and C_L with C'_L . The result, shown in Fig. 3.45(b), indicates a common transfer function for both noise sources.

Straightforward circuit analysis yields

$$A(f) = \frac{v_n}{i_n} = r_{ds1} \parallel r_{ds2} \parallel \frac{1}{j2\pi f C'_L} = \frac{r}{1 + j2\pi f r C'_L},$$

for the transfer function, in which $r \doteq r_{ds1} \parallel r_{ds2}$. The magnitude squared of this transfer function is given by

$$|A(f)|^2 = \frac{r^2}{1 + \left(\frac{f}{f_c}\right)^2}, \quad (3.120)$$

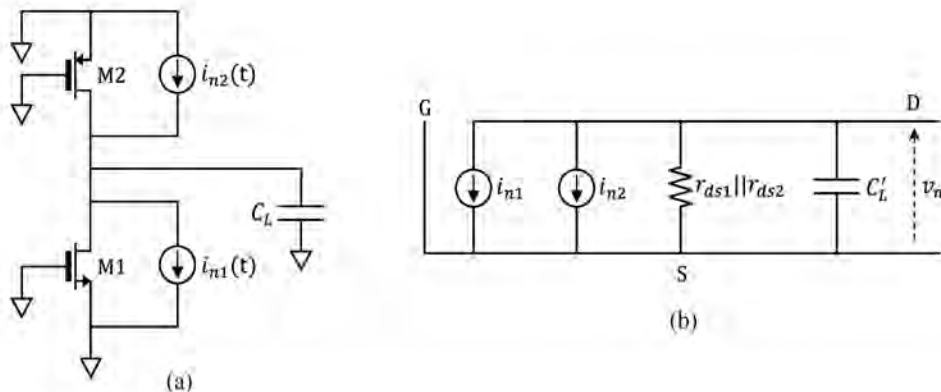


FIGURE 3.45 (a) Noise sources in the current-source-loaded amplifier stage of Example 3.17. (b) Small-signal equivalent circuit for determining the transfer function from noise current to output voltage.

where $f_o \doteq 1/2\pi rC_L'$. Next, we adopt (3.118) and (3.119) for the power spectral densities of thermal and flicker noise components:

$$S_{th1} = \frac{8}{3} E_{th} (g_{m1} + g_{mb1} + g_{ds1}),$$

$$S_{th2} = \frac{8}{3} E_{th} (g_{m2} + g_{mb2} + g_{ds2}),$$

$$S_{f1} = K_1 \frac{W_1}{fL_1^3},$$

$$S_{f2} = K_2 \frac{W_2}{fL_2^3}.$$

Now, we determine the output power spectral density $S_o(f)$ by substituting these four equations together with (3.120) into (2.107), and integrate the latter over f between $f = f_1$ and $f = \infty$ to obtain the following expression for the rms output noise voltage:

$$V_{no}^2 = r^2 f_o (S_{th1} + S_{th2}) \left(\frac{\pi}{2} - \arctan \frac{f_1}{f_o} \right) + r^2 \left(K_1 \frac{W_1}{L_1^3} + K_2 \frac{W_2}{L_2^3} \right) \ln \sqrt{1 + \left(\frac{f_o}{f_1} \right)^2}.$$

Assuming $f_o \gg f_1 = 10$ Hz, this equation is well approximated by

$$V_{no}^2 = \frac{\pi}{2} r^2 f_o (S_{th1} + S_{th2}) + r^2 \left(K_1 \frac{W_1}{L_1^3} + K_2 \frac{W_2}{L_2^3} \right) \ln \left(\frac{f_o}{f_1} \right).$$

For numerical calculation, we adopt $C_L' = 204$ fF, $g_{m1} = 244$ μ A/V, and $r = 340$ k Ω directly from the solution of Example 3.16. We also rerun the operation-point analysis of the input file `ex3-16.sp` to obtain $g_{m2} = 240$ μ A/V, $g_{mb1} = 68.3$ μ A/V, $g_{mb2} = 74.6$ μ A/V in addition to the overdrive voltages $V_{GS1} - V_{TN} = 0.167$ V, $V_{GS2} - V_{TP} = -0.134$ V for which Figs. 3.44(b) and (c) indicate $K_1 = 4.4 \times 10^{-20}$ A² μ m² and $K_2 = 9.9 \times 10^{-21}$ A² μ m². Effective dimensions, calculated from (3.1) and (3.2), are $W_1 = 5.98$ μ m, $L_1 = 0.76$ μ m, $W_2 = 24$ μ m, and $L_2 = 0.74$ μ m. These numbers lead to $f_o = 2.29$ MHz. Considering also $f_1 = 10$ Hz, we arrive at $V_{no} = 2.05$ mV.

Spice verification can be performed with the very same input file `ex3-16.sp` of Example 3.16 after being supplemented with a noise-analysis line

```
.noise v(2) vi dec 100 10 1g
```

V_n , printed with `print onoise_total`, turns out to be 2.06 mV. ▲

3.9 MOSFET as a Switch MOS晶体管用作开关

As discussed in Sections 2.11 and 2.12, external networks of discrete-time amplifiers contain switches. A switch is implemented with either a single n-channel or p-channel MOSFET, or a parallel combination of the two. The latter type of switch is known as *transmission gate*. In this section, we begin with the first-order properties of these two types of switches, and subsequently move to a description of the parasitic *charge-injection* and *clock-feedthrough* effects associated with switching MOSFETs. The reader will recall from the contents of Subsection 2.11.3 that these two effects are responsible for the charge error observed in the dc response of discrete-time amplifier configurations.

3.9.1 Single-Device and Transmission-Gate Switch Properties

单个器件和传输门的开关特性

An NMOS-only switch is depicted in Fig. 3.46(a), where v_i is the input voltage, v_o is the output voltage, and v_G is the gate voltage. All of these voltages are defined with respect to the ground. Also notice the termination of the output port by an equivalent load capacitance C_L representing the switched capacitors and other parasitic capacitances attached to this port.

The state of the switch is controlled primarily by the binary v_G . When the latter is at the negative rail in a split-supply configuration or at the ground in a single-supply

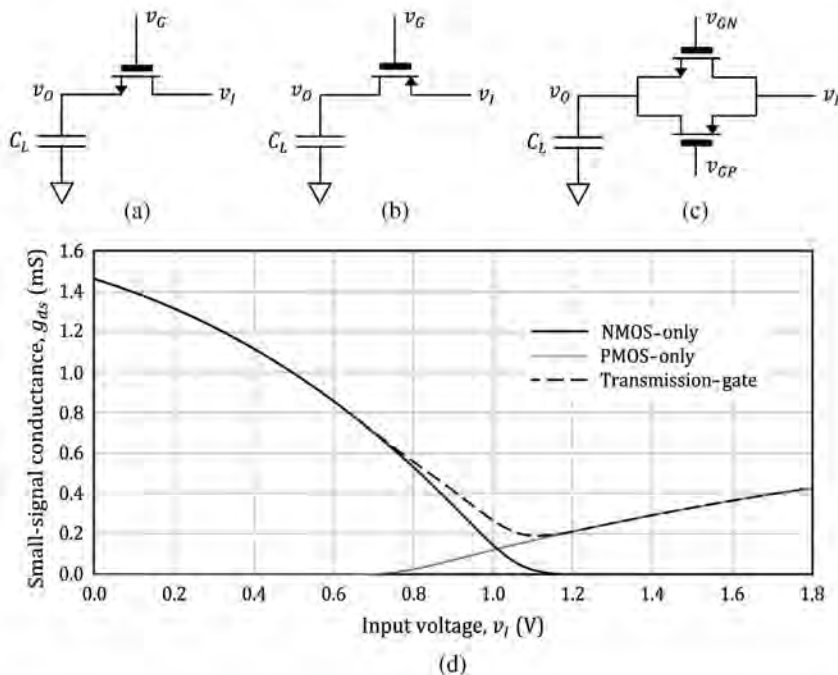


FIGURE 3.46 (a) NMOS-only switch. (b) PMOS-only switch. (c) Transmission-gate switch. (d) Variation of small-signal conductance with input voltage in NMOS, PMOS, and transmission-gate switches. Device dimensions are $W_{DR} = 1 \mu\text{m}$ and $L_{DR} = 0.2 \mu\text{m}$ in all cases.

configuration, the NMOS device remains in cutoff regardless of the input and output voltages. Since conductance virtually vanishes in cutoff, the device acts like an open switch. When v_G is raised to the positive rail V_{DD} , the device exits cutoff and acts like a closed switch with a finite conductance provided that the smaller between v_I and v_O is also smaller than $V_{DD} - V_{TN}$. Otherwise it continues staying in cutoff as implied by (3.8).

For $v_G = V_{DD}$, the typical variation of the dc small-signal conductance g_{ds} of an NMOS switch with the input voltage v_I is displayed in Fig. 3.46(d) by the plot labeled “NMOS only.” This plot has been extracted with a dc-sweep simulation on a standard-threshold device of dimensions $W_{DR} = 1 \mu\text{m}$ and $L_{DR} = 0.2 \mu\text{m}$. It clearly shows how the conductance continuously decreases with an increasing v_I and vanishes beyond $v_I = 1.1 \text{ V}$. This observation indicates a threshold voltage around $V_{TN} = 1.8 - 1.1 = 0.7 \text{ V}$, which is indeed close to what Fig. 3.7 indicates for $L_{DR} = 0.2 \mu\text{m}$ and $V_{SB} = 1.1 \text{ V}$. Note that a high conductance is desirable for a switch because it lowers the time constant of settling. This is why an NMOS switch can offer an acceptable level of dynamic performance only if v_I is confined to the lower half of the rail-to-rail supply-voltage range.

A PMOS switch, shown in Fig. 3.46(b), acts as an open switch for $v_G = V_{DD}$, and is conditionally closed when v_G is lowered to V_{SS} in a split-supply configuration or to the ground in a single-supply configuration. The condition of closure is that the larger between v_I and v_O should be larger than $V_{SS} - V_{TP}$ in a split-supply configuration, or larger than $-V_{TP}$ in a single-supply configuration.

The conductance g_{ds} of a PMOS switch increases with v_I as indicated by the plot extracted with a dc-sweep simulation on a PMOS device of $W_{DR} = 1 \mu\text{m}$ and $L_{DR} = 0.2 \mu\text{m}$ and labeled “PMOS only” in Fig. 3.46(d). Obviously, a reasonable dynamic performance can be displayed by a closed PMOS switch only if v_I happens to be in the upper half of the rail-to-rail supply-voltage range where the conductance is high. Also note that the conductance of a PMOS switch is considerably lower than the conductance of a comparably sized NMOS switch. This is due to the fact that the mobility of holes is approximately one-third of the mobility of electrons. However, this deficiency can be compensated for by widening the PMOS switch.

Considering the fact that NMOS and PMOS switches perform well in complementary ranges of v_I , a full rail-to-rail performance can be achieved by paralleling the two devices as depicted in Fig. 3.46(c). This combination is what we call a *transmission gate*. Obviously, a transmission gate is open when v_{GP} is high and v_{GN} is low. Transposing these states closes the switch. The conductance of a transmission gate is expected to be the sum of NMOS and PMOS conductances, which is indeed what the simulated plot marked “transmission gate” in Fig. 3.46(d) indicates.

3.9.2 Charge Injection and Clock Feedthrough 电荷注入和时钟馈通

For device-level modeling of charge-injection and clock-feedthrough effects, consider an NMOS switch first, and suppose that it opens at an instant $t = t_1$. Its closed state just before this instant is detailed in Fig. 3.47(a), where C_L represents the capacitive load at the output and C_{ov} represents the extrinsic overlap capacitance modeled in Subsection 3.7.3 by (3.92) or (3.93). Notice the instantaneous value $v_I(t_1^-) = V$ of the input voltage and $v_G(t_1^-) = V_{DD}$ of the closed state. If settling is already complete by t_1 , the instantaneous

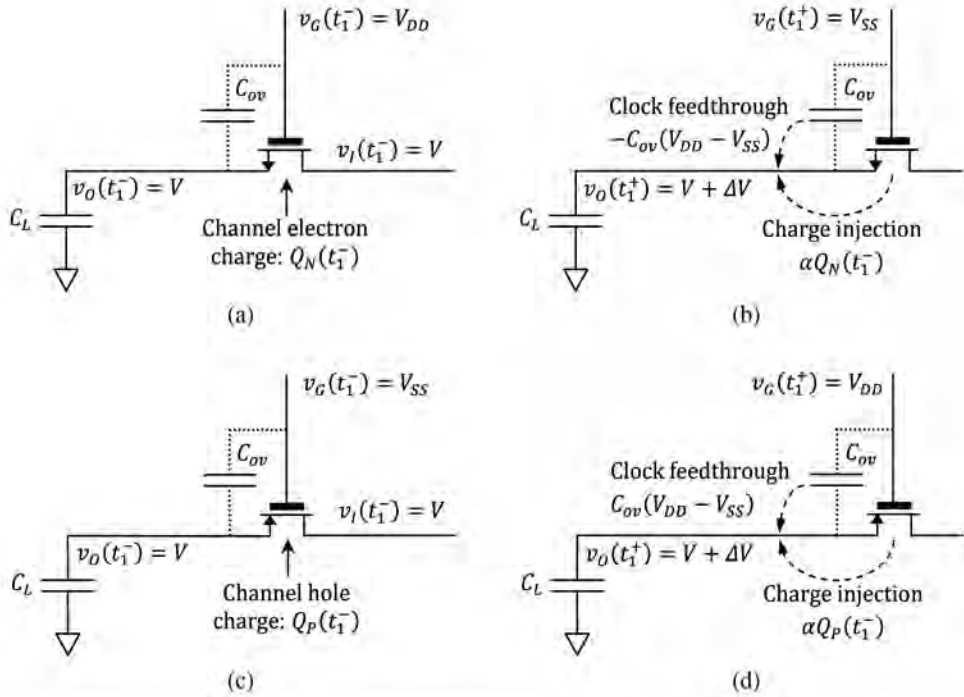


FIGURE 3.47 (a) Closed NMOS switch. (b) Open NMOS switch. (c) Closed PMOS switch. (d) Open PMOS switch.

value of the output voltage equals V too. According to the strong-inversion model, total electron charge associated with the inversion layer of the channel under these conditions is given by

$$Q_N(t_1^-) = -WLC_{ox}(V_{DD} - V - V_{TN}).$$

At $t = t_1$, v_G is lowered to V_{SS} in a split-supply configuration or to ground in a single-supply configuration, and therefore the device is forced into cutoff. As a result of cutoff, $Q_N(t_1^-)$ is transferred to input and output nodes because the inversion layer can no longer be sustained. The part $\alpha Q_N(t_1^-)$ injected to the now floating output node as depicted in Fig. 3.47(b) for $t = t_1^+$ is responsible for what we call the charge-injection effect. It is one of the two components of the charge error Q defined in Subsection 2.11.3. The fraction α depends on various factors including the capacitive load C_L , the transition rate of v_G , and the resistance of the input source, but its value is typically in between 0.25 and 0.75.

Now consider the overlap capacitance C_{ov} . At $t = t_1$, the voltage of the upper plate of this capacitance is lowered from V_{DD} to V_{SS} or to the ground depending on the power-supply configuration. Assuming the general case of a split-supply configuration, and a negligible output voltage perturbation ΔV in comparison with $V_{DD} - V_{SS}$, this change reduces the charge of the upper plate by $C_{ov}(V_{DD} - V_{SS})$. Since an equal amount of charge leaves the lower plate at the output node, the total charge of the output node will change by $-C_{ov}(V_{DD} - V_{SS})$. This is the so-called clock-feedthrough component of the charge error Q .

Combining the contributions of charge injection and clock feedthrough, we can describe the total charge delivered to the output node with

$$Q = \underbrace{-\alpha WLC_{ox}(V_{DD} - V - V_{TN})}_{\text{Charge injection}} \underbrace{-C_{ov}(V_{DD} - V_{SS})}_{\text{Clock feedthrough}} \quad (3.121)$$

Note that, since both components of Q are negative in an NMOS switch, the resulting perturbation ΔV in the output voltage is in the negative direction. Also notice the proportionality of the charge-injection component to channel length L and width W . Clock feedthrough is also proportional to channel width as indicated by (3.92) and (3.93). This is why a small device size is essential for minimizing Q .

A PMOS switch is closed for $V_G = V_{SS}$ and open for $V_G = V_{DD}$. The hole charge residing in its channel prior to opening is represented by $Q_P(t_1^-)$ in Fig. 3.47(c), and is modeled by

$$Q_P(t_1^-) = -WLC_{ox}(V_{SS} - V - V_{TP}).$$

This charge is positive because the overdrive voltage $V_{SS} - V - V_{TP}$ is negative prior to $t = t_1$. Opening the switch injects a fraction $\alpha Q_P(t_1^-)$ of this charge into the output node. The charge contributed to the same node by clock feedthrough amounts to $C_{ov}(V_{DD} - V_{SS})$. Therefore, the total charge received by the output node due to switch opening is described by

$$Q = \underbrace{-\alpha WLC_{ox}(V_{SS} - V - V_{TP})}_{\text{Charge injection}} \underbrace{+C_{ov}(V_{DD} - V_{SS})}_{\text{Clock feedthrough}} \quad (3.122)$$

which, being definitely positive, perturbs the output node in the positive direction. Except for the polarity, however, all of its features are qualitatively similar to those of an NMOS switch.

As pointed out above, charge error is an increasing function of MOSFET channel dimensions. Adopting the minimum permissible dimensions therefore helps minimize this error. If the charge error thus minimized is still unacceptably large, we may prefer supplementing the output node with a *dummy switch* or using a transmission gate instead of a single MOSFET. We now discuss these two techniques of charge-error reduction.

The dummy-switch technique is illustrated in Fig. 3.48 on an NMOS switch. Notice the placement of short-circuited source and drain terminals of a dummy NMOS device MD to the output node of the main switch MM. The clock driving the gate terminal of MD is complementary to the clock driving the gate of MM. Channel length is minimized in both devices but the width of MD is exactly one half of the width of MM. Each of the gate-to-source and gate-to-drain overlap capacitances of MD is therefore exactly one half of the gate-to-output overlap capacitance C_{ov} of MM. However, since both source and drain terminals of MD are connecting the output node, the total gate-to-output overlap

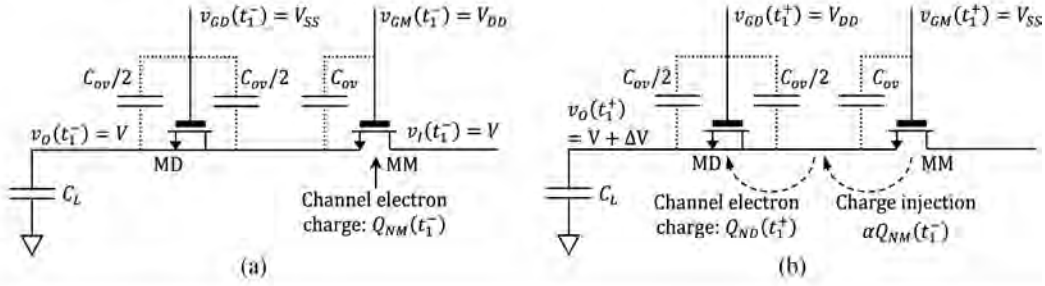


FIGURE 3.48 Dummy switch. (a) Before opening. (b) After opening.

capacitance of this device also equals C_{ov} . Thanks to equal overlap capacitances and complementary clocks, the clock-feedthrough effects of MM and MD cancel out.

Now, consider the state of the MOSFETs at $t = t_1^-$ just prior to switch opening as illustrated in Fig. 3.48(a). Electron charge in the inversion layer of MM is described by

$$Q_{NM}(t_1^-) = -WLC_{ox}(V_{DD} - V - V_{TN}),$$

whereas the cutoff state precludes inversion in MD. Next, consider the state of these devices at $t = t_1^+$ just after switch opening as shown in Fig. 3.48(b). On the one hand, MM enters into cutoff and releases a fraction $\alpha Q_{NM}(t_1^-)$ of its inversion charge to the output node. On the other hand, MD is driven into inversion, and therefore removes a charge

$$Q_{ND}(t_1^+) = -\frac{1}{2}WLC_{ox}(V_{DD} - V - V_{TN})$$

from the same node in order to establish its own inversion layer.⁸ Therefore, the net charge injected into the output node is given by

$$Q = \left(\frac{1}{2} - \alpha\right)WLC_{ox}(V_{DD} - V - V_{TN}). \quad (3.123)$$

Comparing this expression with (3.121), we conclude that the dummy-switch technique not only eliminates clock feedthrough but also reduces charge injection because, as pointed out above, α is usually within the range 0.25 to 0.75.

A transmission gate built with identically sized devices has a first-order immunity against clock feedthrough because devices are controlled by complementary clocks. Its charge-injection performance is similar to a single-NMOS switch for small values of the input voltage and to a single-PMOS switch for large values of the input voltage. However, it performs better than single-device switches for midrange input voltages. To see why, consider a closed transmission gate. For $V < V_{SS} - V_{TP}$, its PMOS is in cutoff with a negligible hole charge in its channel. NMOS, on the other hand, is strongly inverted

⁸ This equation is valid as long as the perturbation caused by charge injection is negligible in comparison with V .

with a negative electron charge. Therefore, switch opening causes only the NMOS to inject charge to the output node. For $V > V_{DD} - V_{TN}$, the roles of NMOS and PMOS are transposed. Now, only the PMOS device contains charge in its channel when the switch is closed. Opening the switch causes part of this positive charge to be injected. If, however, V is in between $V_{SS} - V_{TP}$ and $V_{DD} - V_{TN}$, both devices contain a mobile charge of opposite polarity in their respective channels when the switch is closed. Opening the switch results in positive charge injection from PMOS, and negative charge injection from NMOS. The net charge injected is therefore less than what would be injected by a single device. As a matter of fact, the net charge injected vanishes if V happens to be midway between V_{SS} and V_{DD} . For an analytical proof of this fact, consider a transmission gate with identically sized NMOS and PMOS devices. Also assuming an identical injection factor α for both devices, and summing up the charge-injection components of (3.121) and (3.122), the net charge injected is formulated as

$$Q = -\alpha WLC_{ox}(V_{DD} + V_{SS} - 2V - V_{TN} - V_{TP}),$$

Obviously the charge vanishes for

$$V = \frac{V_{DD} - V_{TN} + V_{SS} - V_{TP}}{2}.$$

Since $V_{TN} + V_{TP}$ is usually very small, the voltage indicated by this equation is indeed close to $(V_{DD} + V_{SS})/2$.

Example 3.18 In this example, we simulate the transient response of the discrete-time amplifier of Example 2.32 with transmission-gate switches.

Work Input file is as follows:

```
filename: ex3-18.sp
.incl bsim.sp
.param vol=0.15 voh=1.65 vdir=5 gm0=1m ro=1meg
.param width=4u
*****ota*****
.subckt ota ninv inv out
gm 0 out cur='gm0*vdir*tanh((v(ninv)-v(inv))/vdir)'
rout out 0 ro
bb 0 out i=0.9/ro
.ends ota
*****transmission gate*****
.subckt tg in out ngate pgate
mn in ngate out 0 cmosn w=width l=0.2u ad={width*0.55e-6}
+ pd={1.1e-6+2*width} as={width*0.55e-6} ps={1.1e-6+2*width}
mp in pgate out 1 cmosp w=width l=0.2u ad={width*0.55e-6}
+ pd={1.1e-6+2*width} as={width*0.55e-6} ps={1.1e-6+2*width}
vbulk 1 0 1,8
.ends
*****amplifier*****
xota 1 2 3 ota
xs1 2 3 7 9 tg
xs2 6 4 7 9 tg
```

```

xs3 6 5 7 9 tg
xs4 4 3 8 10 tg
xs5 5 1 8 10 tg
c1 5 2 2p
c2 4 2 1p
ci 3 0 0.5p
vb 1 0 0.9
vi 6 1 pwl 0 -0.15 74n -0.15 76n 0.15 189n 0.15
+ 191n 0 300n 0 302n -0.15 441n -0.15
vphi1 7 0 pulse 1.8 0 47n 2n 2n 58n 112n
vphi1c 9 0 pulse 0 1.8 47n 2n 2n 58n 112n
vphi2 8 0 pulse 0 1.8 51n 2n 2n 50n 112n
vphi2c 10 0 pulse 1.8 0 51n 2n 2n 50n 112n
.tran 10p 441n
.options trtol=1
.end

```

Notice two differences with respect to the original file `ex2-32.sp`. First is the replacement of the ideal-switch subcircuit with a transmission-gate subcircuit, where the channel width of NMOS and PMOS devices is represented by the parameter `width`. Junction dimensions appearing in element lines are also expressed in terms of this parameter for automatic calculation by the simulator. The second difference is the inclusion of the complementary clocks `vphi1c` and `vphi2c` for driving the PMOS devices.

The reader is invited to examine the differences between (a) the ideal output-voltage waveform obtained with the original file, and (b) the one obtained with the present file for the following values of channel width in micrometers: 0.5, 1, 2, 4, 8, 16, 32, 64, 128, and 256. Which channel width offers the best performance? Try to discover the root causes of inferior performance observed for narrower or wider channels. ▲

3.10 Resistor Design 电阻设计

With the exception of unity-gain buffer, all continuous-time amplifier configurations we studied in Chapter 2 contain resistors in their external network. As will be discussed in Chapter 4, resistors also find use inside opamps, most notably within supplementary circuits that generate a current or voltage reference for biasing purposes. In this section, we elaborate on integrated resistor design in CMOS technology.

3.10.1 Resistor Structures and Resistance Modeling 电路结构和阻抗模型

A resistor can be built in any of the standard semiconductor layers of the CMOS structure or in a dedicated layer if offered as a process option. Inspecting Fig. 3.1, we identify n-well, p^+ , n^+ , and doped poly as standard layers suitable for resistor implementation in a twin-well CMOS technology. Cross-section and top views of basic strip-geometry resistors built in these layers are shown in Figs. 3.49(a) to (d). The resistor structures shown in Figs. 3.49(e) and (f) are built-in dedicated layers as will be described later. Note that in all options, v_A and v_B represent terminal voltages defined with respect to the

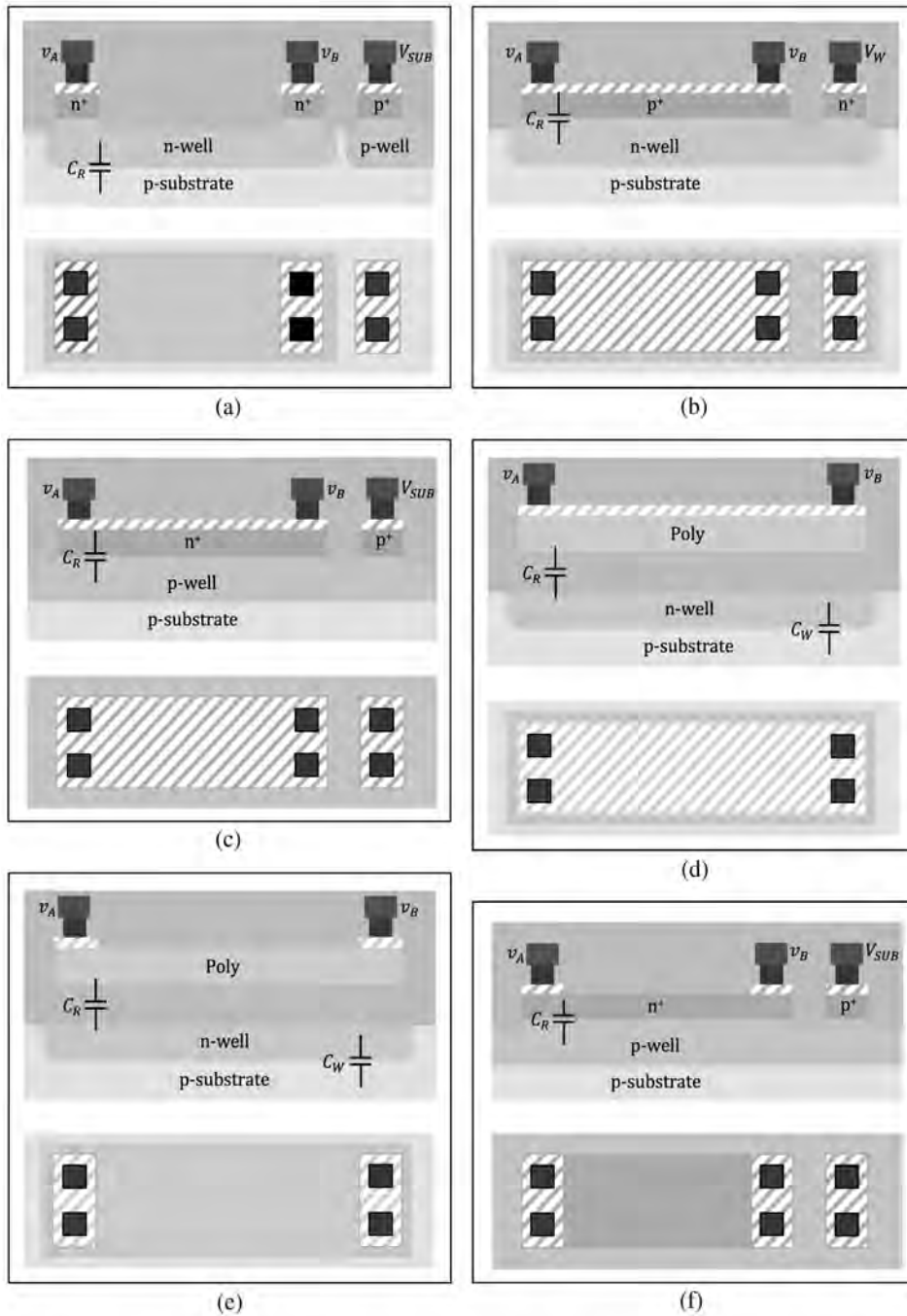


FIGURE 3.49 Resistors built in CMOS semiconductor layers. (a) n -well. (b) p^+ . (c) n^+ . (d) Poly. (e) Poly with silicide block. (f) n^+ with silicide block.

ground. In an n-well or n⁺ resistor, the lower between these two voltages should never be allowed to drop below the substrate voltage V_{SUB} in order not to forward bias the isolating pn junction. By the same token, the higher between the two should not exceed the well voltage V_W in a p⁺ resistor. A poly resistor, not being isolated by a pn junction, is not subject to any such constraint.

The resistance of an integrated resistor is primarily due to the body of the semiconductor layer but the terminal contacts made to the layer also contribute to the resistance. A first-order model including both contributions describes the resistance with

$$R = \underbrace{R_{\square} \frac{L}{W}}_{\text{Body}} + 2 \underbrace{\frac{R_C}{n}}_{\text{Contacts}} \quad (3.124)$$

The contribution of terminal contacts to R is due to a thin interfacial resistive layer between plug material and underlying silicide at contact sites. Its resistance per contact, called *contact resistance*, is represented by R_C in (3.124). In 0.18- μm CMOS, R_C is typically around 10 Ω . Also note that the parameter n appearing in (3.124) represents the number of *primary contacts* paralleled in each of the two heads of the resistance. As indicated in Fig. 3.50, primary contacts are the ones facing the body of the resistor. As also shown in Fig. 3.50, each head may also contain *spare contacts* placed outside the primary contacts as a precaution against the possibility of primary-contact failure. However, their effect on resistance is deemed negligible in proper operation, which is why they do not count toward n . In any case, the contribution of contacts to the resistance of a properly designed resistor should be negligible in comparison with the body component of resistance. The latter, expressed by the first term on the right-hand side of (3.124), is determined by *effective length* L (m), *effective width* W (m), and *sheet resistance* $R_{\square}$ (ohms per square). While the sheet resistance represents the resistance of one square resistive-layer segment of dimensions $W \times W$, the *aspect ratio* L/W corresponds to the number of such squares in between the terminals. Notice in Fig. 3.50 the definition of L between the facing sidewalls of primary-contact plugs.

The process-bias effects explained in Subsection 3.2.1 in the context of MOSFET devices are effective in integrated resistors as well. These effects differentiate W and L from respective drawn values W_{DR} and L_{DR} . Still, it is almost always acceptable to assume

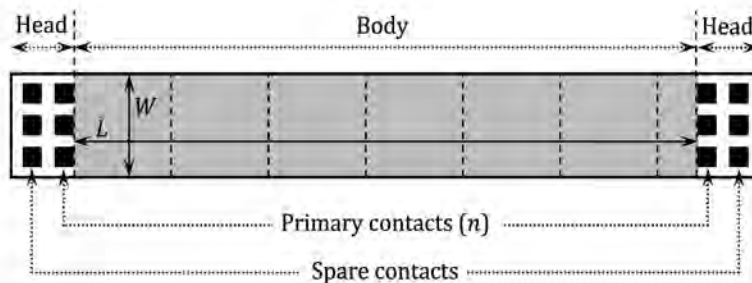


FIGURE 3.50 Length L , width W , and primary and spare contacts in a resistor structure. Gray shading exemplifies a body length of 6.4 squares.

Parameter	Standard layers				Dedicated layers			
	n-well	n ⁺	p ⁺	Poly	n ⁺ with silicide block	p ⁺ with silicide block	Poly with silicide block	High-resistivity poly
$R_{\square}$ (Ω/sq)	950	7	8	8	60	120	300	1100
ΔW (μm)	0.2	-0.08	-0.08	-0.06	0	0	0.08	0.08
$W_{DR(\text{min})}$ (μm)	1.2	0.3	0.3	0.2	0.3	0.3	0.5	0.4

TABLE 3.4 Resistor Layer Data

$L \approx L_{DR}$ because L_{DR} is usually much longer than the process bias associated with the length. This is why we consider only the width bias

$$\Delta W \doteq W_{DR} - W \quad (3.125)$$

in resistance calculations.

Typical values of $R_{\square}$, ΔW , and minimum drawn width $W_{DR(\text{min})}$ in a 0.18- μm CMOS technology are presented in Table 3.4 for all layers of interest. Notice that the sheet-resistance values associated with the standard n⁺, p⁺, and poly layers are relatively low, which is due to the silicidation process applied intentionally for minimizing the parasitic resistances of MOSFET terminals. Implementing, say, a 10-k Ω resistor with any of these three layers would necessitate an aspect ratio in excess of a thousand as one can ascertain from (3.124). Since the minimum of W is technologically limited, such a large aspect ratio implies a very long resistor geometry, hence a large footprint and a large parasitic capacitance. This is why these layers are suitable only for implementing very low resistances. Among the standard layers, only n-well has a sufficiently high sheet resistance for implementing a high-resistance resistor in a relatively small area but its nonlinearity is also high. Obviously, none of the resistor structures shown in Figs. 3.49(a) to (d) is close to ideal.

Two options in most CMOS technologies offer dedicated resistor structures of higher utility. One of the options, called *silicide block*, enables us to turn n⁺, p⁺, or doped poly into a dedicated layer by locally blocking the silicide formation. As listed in Table 3.4, the sheet resistance of an unsilicided layer is an order of magnitude higher than its silicided counterpart. Thanks to this property, a high-resistance resistor can be implemented in a much smaller area. The cross-sectional and top views of such poly and n⁺ resistors are shown in Figs. 3.49(e) and (f). Notice the retention of silicide under the contacts in order to avoid rectification due to Schottky-barrier formation. The second dedicated-layer option is known as *high-resistivity poly*. It relies on the undoped body of a second layer of poly introduced specifically for resistor and capacitor integration. Its geometry is similar to the doped poly resistor shown in Fig. 3.49(e) but, as cited in Table 3.4, it offers a much higher sheet resistance.

The resistance of an integrated resistor is not perfectly linear because the sheet resistance somewhat depends on the applied voltage. This dependence has several different root causes. One, observed in junction-isolated resistors, is the modulation of the effective thickness of the resistor by the voltage-dependent junction depletion region. This effect is more pronounced in layers of relatively lighter doping, which is why n-well resistors are the most nonlinear among all.

Another effect that causes nonlinearity by making the sheet resistance an increasing function of voltage in all types of resistors is the degradation of carrier mobility by the longitudinal electric field. Since this electric field is well approximated by $|v_A - v_B|/L$, the effect is obviously most pronounced in short resistors operating with a high voltage.

The unique grainy morphology of polysilicon is yet another cause for nonlinearity. Being field dependent, this type of nonlinearity is also significant in short poly resistors operating with a high voltage.

Finally, the variation of sheet resistance with temperature is an indirect cause for nonlinearity because the temperature depends on power dissipation hence voltage. This type of nonlinearity is most pronounced in poly resistors whose oxide isolation has a very high-thermal resistance. Still, however, poly resistors are usually the best in overall linearity.

The temperature sensitivity of resistance in junction-isolated resistors is due to mobility degradation and has a positive value. In poly resistors a negative temperature sensitivity may be observed due to additional conduction mechanisms involved at grain boundaries. The effect of voltage and temperature on resistance is conventionally described by generalizing the body component of (3.124) into the polynomial form

$$R = R_{\square} \frac{L}{W} (1 + \text{VCR1} \times v + \text{VCR2} \times v^2) [1 + \text{TCR1} \times (T - T_0) + \text{TCR2} \times (T - T_0)^2], \quad (3.126)$$

where VCR1 are VCR2 voltage coefficients, TCR1 and TCR2 are temperature coefficients, $v \equiv |v_A - v_B|$ is the absolute value of the voltage across the resistor, and T_0 is the temperature at which resistance parameters are characterized. Most technology providers adopt $T_0 = 25^\circ\text{C}$. The model described by (3.126) is reasonably accurate for temperature sensitivity but not quite so for nonlinearity. A curious reader may find more accurate variants in most PDKs and related literature. Also note that TCR1 is around 3000 ppm/C in standard layers and 1000 ppm/C in dedicated layers.⁹ In some technologies, poly layers may exhibit a negative TCR1. TCR2 is typically around 0.5 to 1 ppm/C² in most layers except for high-resistivity poly which usually exhibits a slightly higher value.

Dielectric isolation by means of a junction-depletion region or by an oxide film gives rise to a parasitic capacitance between a resistor and the underlying semiconductor region as represented by C_R in Fig. 3.49. The simplest modeling expression of this capacitance is

$$C_R = C_A A + C_F P, \quad (3.127)$$

where A and P are the area and perimeter of the resistor, and C_A and C_F are two coefficients. Note that A and P should include not only the body of the resistor but also the heads. The first term on the right-hand side of (3.127) represents the areal component. The second term is the component associated with the fringing field along the perimeter. Some PDKs may utilize more sophisticated variants of (3.127) involving voltage-dependent descriptions for C_A and C_F or a distributed capacitance.

⁹“ppm” stands for *parts per million*, that is, 10^{-6} .

Poly resistors usually exhibit the smallest C_A and C_p thanks to the thick oxide used for their isolation. Parasitic capacitance can be further reduced in poly resistors by placing a floating well on the underlying silicon surface as depicted in Figs. 3.49(d) and (e). The series combination of C_R and the well-to-substrate capacitance C_W reduces the effective capacitance to $C_R C_W / (C_R + C_W)$. Finally, note $C_A = 100 \text{ aF}/\mu\text{m}^2$ and $C_F = 40 \text{ aF}/\mu\text{m}$ as the typical coefficient values for dedicated poly resistors in $0.18\text{-}\mu\text{m}$ CMOS technologies.

Example 3.19 Ignoring the contact resistances and relying on the data of Table 3.4, determine the necessary length of a $20\text{-k}\Omega$ resistor of minimum width for the following layers: n^+ , n^+ with silicide block, p^+ with silicide block, poly with silicide block, and high-resistivity poly.

Work

- n^+ : $W = 0.38 \mu\text{m}$ from (3.125), and $L = 1085 \mu\text{m}$ from (3.124).
- n^+ with silicide block: $W_{DR} = 0.3 \mu\text{m}$ from (3.125), and $L_{DR} = 100 \mu\text{m}$ from (3.124).
- p^+ with silicide block: $W_{DR} = 0.3 \mu\text{m}$ from (3.125), and $L_{DR} = 50 \mu\text{m}$ from (3.124).
- Poly with silicide block: $W_{DR} = 0.42 \mu\text{m}$ from (3.125), and $L_{DR} = 28 \mu\text{m}$ from (3.124).
- High-resistivity poly: $W_{DR} = 0.32 \mu\text{m}$ from (3.125), and $L_{DR} = 5.82 \mu\text{m}$ from (3.124). ▲

3.10.2 Design Techniques for Accuracy and Precision

准确度和精确度设计技术

Electrical design of an integrated resistor is completed by determining W_{DR} and L_{DR} after selecting a semiconductor layer that can deliver the specified resistance with acceptable nonlinearity, temperature sensitivity, and parasitic capacitance. Design is based on the mean values of $R_{\square}$, ΔW , and R_C as characterized by the technology vendor over the past runs of the target technology and declared in the PDK as typical values. However, none of the fabricated copies of the resistor will exhibit the target resistance exactly. Their resistance will be distributed randomly around a mean value which itself will be offset from what has been targeted in design. As explained in Subsection 2.4.2, these imperfections will contribute to the imprecision and inaccuracy of any closed-loop performance metric that depends on the resistance of this resistor. Fortunately, the resulting uncertainty can be drastically reduced if we base the metric on the ratio of the resistances of two properly laid-out resistors. A ratio-based metric will also be highly insensitive to temperature. The closed-loop gain of a continuous-time inverting or noninverting amplifier is probably the best-known beneficiary of this practice. Any two resistors intended for accurately and precisely implementing a specific resistance ratio are called *ratioed resistors*. *Ratioing* is actually a very general design practice applied to capacitors and MOSFETs as well. These additional applications will be covered in Subsections 3.11.1 and 4.2.2.

The accuracy of the ratio of two resistances is measured by the difference between the mean and target values of the ratio. Of course, the lesser the difference, the better the accuracy. The mean value deviates from the target value mainly due to (a) imperfect repeatability of fabrication processes; (b) on-chip gradients in resistance parameters,

temperature, and stress; and (c) proximity effects in patterning. Number rounding in design also contributes to inaccuracy. The specific physical-design techniques applied for neutralizing these root causes of inaccuracy in ratioed resistors are *segmentation*, *common-centroid placement*, and *dummy termination*. How these techniques work is described in this subsection.

The precision of the ratio of two resistances is measured by the extent of the distribution of the ratio among fabricated copies. Obviously, the tighter the distribution, the higher the precision. The distribution is mainly due to local random microscopic fluctuations observed in such morphological variables as doping concentration, polysilicon grain size, and edge roughness. As will be discussed in this subsection, the effect of such fluctuations can be reduced only by enlarging the size of ratioed resistors.

Segmentation into Matching Units

Before describing the technique of segmentation, let us go through a hypothetical ratioed-resistor design effort without applying any particular technique for reducing the inaccuracy. Consider the resistor pair whose layout is shown Fig. 3.51(a), and suppose that we intend to implement a ratio $R_A/R_B = 1.5$ with unsilicided poly resistors of $R_A = 15 \text{ k}\Omega$ and $R_B = 10 \text{ k}\Omega$. Aiming for a regular and compact layout, we

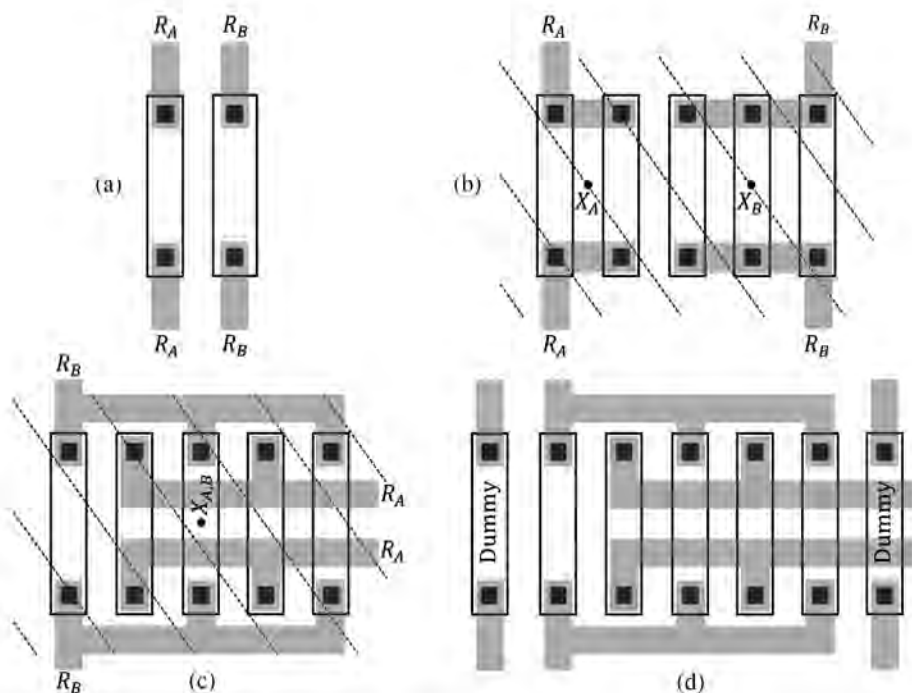


FIGURE 3.51 Resistor ratioing. (a) Ratioing by width adjustment. (b) Ratioing by segmentation into matching unit resistors, where dashed lines are iso-resistivity contours, X_A is the centroid of R_A , X_B is the centroid of R_B . (c) Matching by common-centroid interdigitation where $X_{A,B}$ represents the common centroid. (d) Dummy termination.

decide to adopt equal lengths for the two resistors and the minimum possible drawn-width $W_{DR(A)} = 0.50 \mu\text{m}$ for R_A . These numbers, together with $R_C = 10 \Omega$, $n = 1$, $R_{\square} = 300 \Omega/\text{sq}$, and $\Delta W = 0.08 \mu\text{m}$, as cited in Table 3.4, yield $L_A = 20.97 \mu\text{m}$ from (3.124) and (3.125) for R_A . Adopting the same length for R_B , we calculate the drawn width of this resistor from (3.124) and (3.125) as $W_{DR(B)} = 0.71 \mu\text{m}$. Notice that all dimensions involved in these calculations are rounded to the nearest hundredth of a micrometer because this is the minimum grid size mandated by the PDK for physical design. If we now calculate the resistance ratio with these rounded numbers, we will end up with $R_A/R_B = 1.4990$, hence a -0.067% systematic deviation with respect to the target value. Therefore, the ratioed resistor layout depicted in Fig. 3.51(a) suffers from a small inaccuracy even before being taped out for fabrication!

Let us now consider the inaccuracies inflicted on the resistance ratio of the layout of Fig. 3.51(a) by the imperfect repeatability of fabrication processes. First, imagine a fabrication run in which the mean value $\overline{R_{\square}}$ of sheet resistance happens to be 15% smaller than its typical value $300 \Omega/\text{sq}$, while all other process parameters remain at their respective typical values cited in Table 3.4. This results in approximately -15% deviation in the mean values of the two resistances but the mean value of their ratio shifts only by -0.078% including the deviation due to number rounding. The accuracy of the resistance ratio is obviously far better than what each resistance individually offers.

Next, consider a fabrication run in which the mean value $\overline{\Delta W}$ of process bias in fabricated resistors happens to be $0.05 \mu\text{m}$ instead of the typical $0.08 \mu\text{m}$ considered in design. Assuming typical values for all other process parameters, and recalculating the deviation from target value in each of the two resistances, we arrive at -6.66% for $\overline{R_A}$, -4.48% for $\overline{R_B}$, and -2.28% for the ratio of the two. Although the ratio of two resistances turns out to be still more accurate than an individual resistance, the extent of improvement is nowhere near what we have observed in the previous case of sheet-resistance deviation. Ratioing alone is obviously good for accuracy but it is not perfect.

Fortunately, the inaccuracy emanating from number rounding or imperfect repeatability of fabrication processes can be eliminated almost entirely in ratioed design by segmenting both resistors into physically *matching* unit resistors connected in series or in parallel. A segmented version of the resistors of Fig. 3.51(a) is shown in part-b of the same figure. Notice the use of two parallel unit resistors for R_A , and three parallel unit resistors for R_B . The mean value of the resistance ratio is therefore related to the mean value $\overline{R_U}$ of a unit resistor by

$$\overline{R_A/R_B} = \frac{(1/2) \times \overline{R_U}}{(1/3) \times \overline{R_U}},$$

which equals 1.5 regardless of how $\overline{R_U}$ is affected by number rounding in design or by imperfections of process repeatability.

Interdigitated Common-Centroid Placement

Segmentation is effective on those inaccuracies arising from the imperfect repeatability of fabrication processes but it is not a cure for inaccuracies that are caused by on-chip

nonrandom spatial variations in resistance parameters. Such nonrandom variations are characterized by a common gradient sustained along the same direction in all fabricated copies of the chip. They may be due to spatial nonuniformities in fabrication processes, in operation temperature, or in mechanical stress.¹⁰

Now, suppose that a resistance parameter, say $R_{\square}$, varies across the segmented pair of resistors R_A and R_B as depicted in Fig. 3.51(b) with dashed iso-resistivity contours. As long as the spatial variation is linear, the resistance of each resistor is still described by (3.124) provided that the value of sheet resistance at the centroid is adopted for $R_{\square}$.¹¹ This leads us to the conclusion that if the centroids X_A and X_B of the two resistors are placed on the same constant-resistivity contour, the target ratio $R_A/R_B = 1.5$ will still be accurately achieved despite the spatial variation of sheet resistance. However, it is virtually impossible to predict the pattern of on-chip spatial variations. A radical solution to this problem is to lay out the ratioed resistors with coincident centroids. This technique, called *common-centroid placement*, is exemplified in Fig. 3.51(c) where unit resistors are interdigitated in a pattern B A B A B. Another possible common-centroid pattern is A B B B A but this one is less preferable because interdigitation is not as dispersed as it is in the previous one. Dispersion enables common-centroid placement to function as intended even in the case of mildly nonlinear spatial variations. Finally, note that any type of interdigitation is useful for countering the effect of systematic spatial variations of resistance parameters even if a common centroid cannot be established. Common-centroid placement is the ultimately effective form of interdigitation.

Dummy Termination

The process-bias parameter ΔW is affected by the nonuniformity of the etching processes applied in fabrication. Unfortunately, these processes are somewhat sensitive to the proximity of the structural features neighboring a resistor. Consider, for example, the common-centroid poly resistors depicted in Fig. 3.51(c). During fabrication, the outside facing edges of the leftmost and rightmost B units are usually etched faster than all other edges of both units. This leads to a larger effective ΔW hence a narrower W in the leftmost and rightmost units. The resulting resistance ratio will deviate from the target 1.5 despite the perfectly interdigitated common-centroid placement. This problem can be solved by placing dummy features of the same type of resistive layer alongside the outermost resistor segments as shown in Fig. 3.51(d). Being connected to ground or to a dc voltage, these dummies equalize the effect of proximity on etching for all segments without electrically interfering with the operation of the two ratioed resistors.

Sizing for Precision

Distribution of the resistance ratio among the fabricated copies of ratioed resistors leads to imprecision. The main cause of imprecision is the random spatial fluctuation of morphological variables of resistance in microscopic scale. For example, sidewall irregularities lead to a fluctuation in ΔW , whereas irregularities in layer thickness or in doping

¹⁰ Semiconductor resistors are sensitive to mechanical stress due to their piezoelectric property. Chips are permanently stressed in a nonuniform fashion during packaging.

¹¹ Centroid is defined as the center of mass of an object.

concentration lead to a fluctuation in $R_{\square}$. The resulting distribution is usually Gaussian with a relative standard deviation described by

$$\frac{\sigma[R]}{R} = \frac{K_R}{\sqrt{LW}}, \quad (3.128)$$

where the so-called Pelgrom coefficient K_R is typically in between 0.04 and 0.15 μm for poly resistors in a typical 0.18- μm CMOS technology.[8] According to this expression, distribution can be tightened by elongating and/or widening the resistor.

Example 3.20 Size the 30-k Ω unsilicided-poly unit resistor used in the ratioed-resistor configuration of Fig. 3.51(d) so that the resistance ratio R_A/R_B does not deviate more than 1% of the target value 1.5 in at least 99.9997% of the fabricated copies. Assume $K_R = 0.04 \mu\text{m}$.

Work Since inaccuracy is practically eliminated by the dummy-terminated common-centroid layout, the specified uncertainty can be attributed solely to imprecision. Since 99.9997% corresponds to 4.5σ , the specification implies

$$\frac{\sigma[R_A/R_B]}{R_A/R_B} \leq \frac{0.01}{4.5} = 2.22 \times 10^{-3},$$

which, considering R_A and R_B as two uncorrelated random variables in (2.28), can be expressed as

$$\frac{\sigma[R_A/R_B]}{R_A/R_B} = \sqrt{\frac{\sigma^2[R_A]}{R_A^2} + \frac{\sigma^2[R_B]}{R_B^2}} \leq 2.22 \times 10^{-3}. \quad (3.129)$$

Since R_A has two parallel unit resistors and R_B has three, the following expressions hold:

$$R_A = \frac{1}{\frac{1}{R_{U1}} + \frac{1}{R_{U2}}},$$

and

$$R_B = \frac{1}{\frac{1}{R_{U3}} + \frac{1}{R_{U4}} + \frac{1}{R_{U5}}}.$$

where R_{U_i} , $i = 1, \dots, 5$, are five independent random variables representing unit resistances of nominal value R_U and standard deviation $\sigma[R_U]$. Applying (2.28) once more, we arrive at

$$\frac{\sigma^2[R_A]}{R_A^2} = \frac{1}{2} \frac{\sigma^2[R_U]}{R_U^2},$$

and

$$\frac{\sigma^2[R_B]}{R_B^2} = \frac{1}{3} \frac{\sigma^2[R_U]}{R_U^2}.$$

Substitution of these two into (3.129) yields

$$\frac{\sigma[R_A/R_B]}{R_A/R_B} = 0.913 \frac{\sigma[R_U]}{R_U} \leq 2.22 \times 10^{-3},$$

hence

$$\frac{\sigma[R_U]}{R_U} \leq 2.43 \times 10^{-3}.$$

Using this result together with $K_R = 0.04 \mu\text{m}$ in (3.128), we obtain

$$WL \geq 271 \mu\text{m}^2 \quad (3.130)$$

for the effective area of a unit resistor. W and L are interrelated also through (3.124). Ignoring the resistance of the contacts, and adopting $R_U = 30 \text{ k}\Omega$ from design specifications and $R_{\square} = 300 \Omega/\text{sq}$ from Table 3.4, this equation yields $L/W = 100$, which, together with (3.130), yield $W = 1.65 \mu\text{m}$ and $L = 165 \mu\text{m}$. According to (3.125) and Table 3.4, drawn width should be set to $W_{DR} = 1.73 \mu\text{m}$. ▲

3.10.3 MOSFET as a Resistor MOS晶体管用作电阻

The drain-source port of a MOSFET operating in nonsaturation can be utilized as a resistor. Considering the strong-inversion characteristic equation (3.20) of the drain-source port of an NMOS device, we can express the small-signal port resistance $r_{ds} \triangleq 1/(\partial i_D/\partial v_{DS})$ with the equation

$$r_{ds} = \underbrace{\frac{1}{\mu_N C_{OX}(V_{GS} - V_{TN} - A_N V_{DS})}}_{R_{F1}} \frac{L}{W} \quad (\text{n-channel in nonsaturation}). \quad (3.131)$$

For a PMOS device, port characteristic in nonsaturation is described by (3.48), and small-signal resistance is defined by $r_{ds} \triangleq -1/(\partial i_D/\partial v_{DS})$. These two yield

$$r_{ds} = \underbrace{\frac{1}{\mu_P C_{OX} [A_P V_{DS} - (V_{GS} - V_{TP})]}}_{R_{\square}} \frac{L}{W} \quad (\text{p-channel in nonsaturation}). \quad (3.132)$$

Notice that these resistance expressions are in the same form as the body resistance component of (3.124). However, $R_{\square}$ is not a constant but a function of the overdrive and drain-source voltages at the operation point. Dependence on the overdrive voltage makes the resistance electrically controllable, which is a useful property exploited widely in tunable continuous-time MOSFET-C filters. Variation with V_{DS} , on the other hand, is undesirable because it implies a significant nonlinearity. Although this problem can be alleviated by deploying fully-differential closed-loop configurations and certain composite MOSFET techniques, it is still virtually impossible for a MOSFET resistor to match the level of linearity available from passive resistors.

Example 3.21 Shown in Fig. 3.52(a) is a low-pass filter in the form of leaky integrator. Its small-signal transfer function as determined by following the extraction technique described in Subsection 2.10.2 is given by

$$A_{cl} \triangleq \frac{v_o}{v_i} = -\frac{R_2}{R_1} \frac{1}{1 + sCR_2}, \quad (3.133)$$

where R_1 represents R_{A1} and R_{B1} , R_2 represents R_{A2} and R_{B2} , and C represents C_A and C_B . Notice the dc gain $A_{cl0} = -R_2/R_1$ and the closed-loop bandwidth $f_{cl} = 1/2\pi CR_2$. A MOSFET-C implementation is to be designed in accordance with the following specification set:

1. Input-signal swing: $v_{i(\min)} = -0.2$ V, $v_{i(\max)} = 0.2$ V.
2. Closed-loop dc gain: $A_{cl0} = 2$ V/V.
3. Closed-loop bandwidth f_{cl} at the operation point: Tunable in between 6 kHz and 12 kHz by means of a dc voltage in between 0.1 and 1.7 V. Exact limits of the control-voltage range are to be determined.
4. Capacitors: 25 pF each.
5. Opamp: opmacro2.sp specified by $V_{DD} = 1.8$ V, $V_{SS} = 0$, $V_{OH} = 1.65$ V, $V_{OL} = 0.15$ V, $A_{dijf0} = 10000$ V/V, $V_{DIR} = 0.5$ V, $f_{GBW} = 12$ MHz, $f_{nd} = 10$ GHz.
6. Common-mode voltages $V_{ICM} = V_{OCM}$: Value to be determined.

Work Simulation schematic for an NMOS implementation is shown in Fig. 3.52(b), where all MOSFETs have identical geometries and are controlled by the same gate

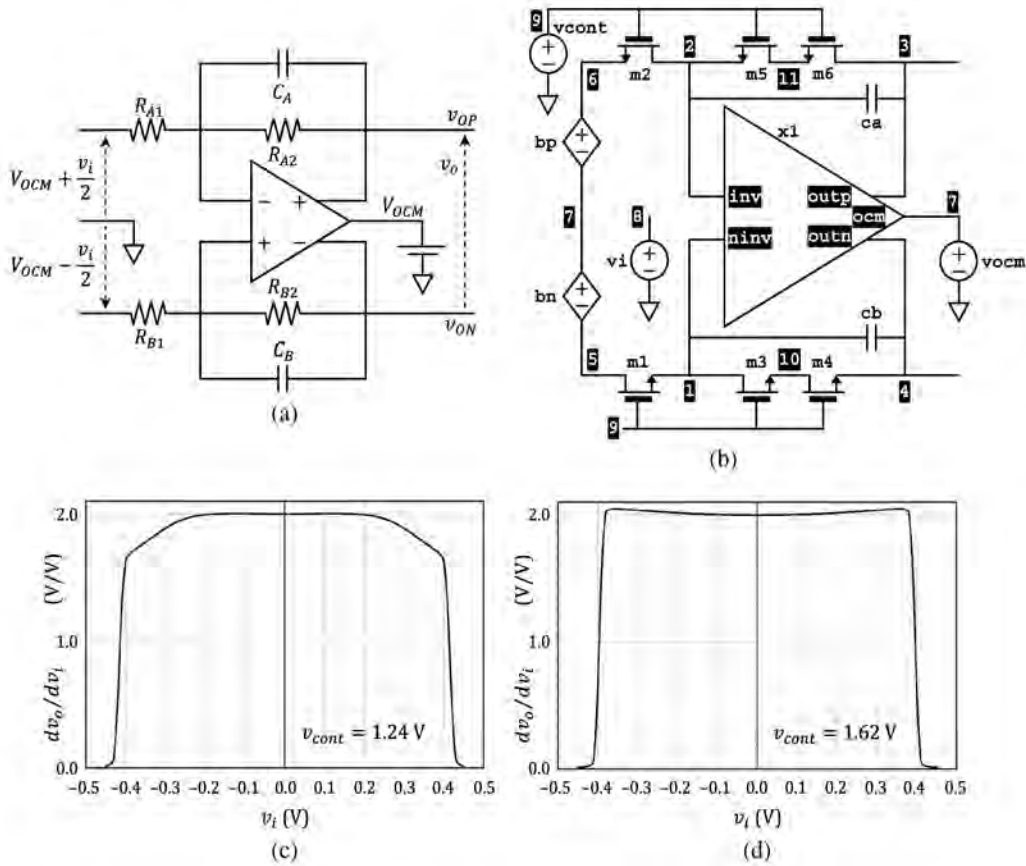


FIGURE 3.52 Low-pass filter of Example 3.21. (a) Circuit schematic. (b) Simulation schematic. (c) DC gain as a function of input signal for $v_{cont} = 1.24$ V. (d) DC gain as a function of input signal for $v_{cont} = 1.62$ V.

voltage V_{cont} . R_2 is implemented by the series combination of unit devices M3 and M4 in the lower half, and M5 and M6 in the upper half. As represented by M1 and M2, one unit device is used for implementing R_1 in each half. The resulting setting $R_2/R_1 = 2$ automatically satisfies item 2 of the specification set.

Items 1 and 2 of the specification set imply $V_{O(min)} = V_{OCM} - 0.2$ and $V_{O(max)} = V_{OCM} + 0.2$ for the limits of single-ended output swing. Considering the necessary conditions $V_{O(min)} > V_{OL}$ and $V_{O(max)} < V_{OH}$ in addition to the single-ended output range limits $V_{OL} = 0.15$ V and $V_{OH} = 1.65$ V of the opamp, we arrive at the following constraint on V_{OCM} :

$$0.35 \text{ V} \leq V_{OCM} \leq 1.45 \text{ V}. \quad (3.134)$$

Another design constraint which is implicit in any MOSFET-resistor application is imposed by the condition of nonsaturation. This condition must be satisfied throughout the entire range of signal swing in order to ensure a reasonable linearity. For the NMOS devices used in this design, it necessitates $V_{DS(max)} < V_{DSsat(min)}$, where $V_{DS(max)}$ is the

largest drain-source voltage and $V_{DSsat(min)}$ is the smallest drain-source saturation voltage experienced by a unit device. In order to identify the unit device that is most unfavorably positioned in this respect, consider the extreme case of $v_i = -0.2$ V. In the upper half of the schematic in Fig. 3.52(b), the node voltages are expected to be $v(6) = V_{OCM} - 0.1$, $v(2) \approx V_{OCM}$, $v(3) \approx V_{OCM} + 0.2$, and $v(11) \approx V_{OCM} + 0.1$. In the lower half, we expect $v(5) = V_{OCM} + 0.1$, $v(1) \approx V_{OCM}$, $v(4) \approx V_{OCM} - 0.2$, and $v(10) \approx V_{OCM} - 0.1$. Obviously, all of the devices operate with approximately the same maximum drain-source voltage $V_{DS(max)} = 0.1$ V. As to $V_{DSsat(min)}$, we need to identify the device that operates with the lowest overdrive voltage and therefore the highest source voltage. M6 happens to be the one with a source voltage $V_{OCM} + 0.1$. Since the overdrive voltage of this device is absolutely minimized when its gate voltage is lowered to $V_{cont(min)}$, the nonsaturation condition finally turns into

$$\underbrace{\frac{V_{cont(min)} - (V_{OCM} + 0.1) - V_{TN}|_{v_s=V_{OCM}+0.1}}{A_N}}_{V_{DSsat(min)}} \geq \underbrace{0.1}_{V_{DS(max)}} \quad (3.135)$$

The factor-of-two controllability of the bandwidth as required in item 3 of the specification set imposes yet another design constraint. Note that the bandwidth is inversely proportional to R_2 , and is specified for the operation point. Since R_2 is described generally by (3.131), in which, the operation point condition $v_i = 0$ of the configuration implies $V_{DS} = 0$ and $V_S = V_{OCM}$, this constraint can be expressed as

$$\frac{V_{cont(max)} - V_{OCM} - V_{TN}|_{v_s=V_{OCM}}}{V_{cont(min)} - V_{OCM} - V_{TN}|_{v_s=V_{OCM}}} = 2. \quad (3.136)$$

It is worth noting that this expression is not very accurate because it does not take into account the variation of mobility with gate voltage, which may have a significant effect at large overdrive levels.

The design constraints defined so far enable us to determine V_{OCM} , $V_{cont(min)}$, and $V_{cont(max)}$. Assuming V_{OCM} as an independent design variable, we first calculate $V_{cont(min)}$ from (3.135) after adding 0.1 V to the right-hand side as a safety margin. Next, we determine $V_{cont(max)}$ from (3.136). $V_{TN}|_{v_s=V_{OCM}+0.1}$ appearing in (3.135) and $V_{TN}|_{v_s=V_{OCM}}$ appearing in (3.136) are determined from Fig. 3.7. The outcome of these calculations are presented in Table 3.5. Note that the minimum entry for V_{OCM} is 0.35 V as limited by (3.134). In contrast, the 0.5-V maximum entry for the same design variable is much lower than what (3.134) indicates as the upper limit. The reason is that, for this value of V_{OCM} , the upper limit 1.7 V imposed on $V_{cont(max)}$ in item 3 of the specification set is already reached. Obviously, the viable range of V_{OCM} is very narrow. The midpoint $V_{OCM} = 0.40$ V of this range is a reasonable choice. The corresponding limits of the control voltage are $V_{cont(min)} = 1.24$ V and $V_{cont(max)} = 1.59$ V.

The final step of design is device sizing in accordance with the bandwidth specification cited in item 3 of the specification set. Since item 4 prescribes 25 pF for the capacitance,

V_{CM} (V)	$V_{TN} _{V_S=V_{CM}+0.1}$ (V) from Fig. 3.7	$V_{cont(min)}$ (V) from (3.135) + 0.1	$V_{TN} _{V_S=V_{CM}}$ (V) from Fig. 3.7	$V_{cont(max)}$ (V) from (3.136)
0.35	0.50	1.18	0.48	1.53
0.40	0.51	1.24	0.49	1.59
0.45	0.53	1.31	0.50	1.67
0.50	0.54	1.37	0.51	1.73

TABLE 3.5 Calculated Data of Example 3.21

R_2 should be around 1.06 M Ω in order to realize the specified 6-kHz lower limit of the bandwidth. This implies $r_{ds} = 530$ k Ω for each unit device operating with a gate voltage $V_G = V_{cont(min)} = 1.24$ V. Considering also $v_s = V_{OCM} = 0.4$ V for the source voltage at the operation point, the aspect ratio W/L of the MOSFET can be calculated from (3.131) after determining the mobility from (3.17). However, sizing by iterative simulation instead of calculation will be more efficient because an input file has to be written for verification anyway. Shown below is the input file `ex3-21.sp` corresponding to the schematic shown in Fig. 3.52(b).

```
filename: ex3-21.sp
.param vdd=1.8 vss=0 voh=1.65 vol=0.15 adif0=10000 vdir=0.25
+ fgbw=10meg fnd=10g
.incl opmacro2.sp
.incl bsim.sp
x1 1 2 3 4 7 fdopamp
ca 2 3 25p
cb 1 4 25p
m1 5 9 1 0 cmosn w=0.5u l=26u
m2 6 9 2 0 cmosn w=0.5u l=26u
m3 1 9 10 0 cmosn w=0.5u l=26u
m4 10 9 4 0 cmosn w=0.5u l=26u
m5 2 9 11 0 cmosn w=0.5u l=26u
m6 11 9 3 0 cmosn w=0.5u l=26u
vcm 7 0 0.40
bp 6 7 v=v(8)/2
bn 7 5 v=v(8)/2
vcont 9 0 1.24 ; 1.62
vi 8 0 dc 0 ac 200m
.ac dec 100 1 100k
.measure ac fcl when vm(4)=141.42m
*.dc vi -450m 450m 10m
.end
```

Considering the rather large target value of r_{ds} , the aspect ratio W/L is likely to be less than unity. It is therefore reasonable to set a narrow width and to iterate an ac analysis by varying the channel length until the corner frequency meets the specification. Notice the selection of $W_{DR} = 0.5$ μm and the inclusion of a measurement line “`.measure`” in

the input file. Since the ac magnitude of the differential input signal source v_i is set to 0.2 V, and the low-frequency gain from this source to each single-ended output equals 1 V/V, the ac magnitude of a single-ended output signal at the corner frequency must equal $0.2(\text{V}) \times 1(\text{V/V})/\sqrt{2} = 0.14142 \text{ V}$. This is why the measurement line instructs the simulator to measure the frequency at which the ac magnitude $v_m(4)$ of one of the single-ended outputs is exactly this much. Simulator puts out the result as $fc1$ after running the ac analysis. Iterating with a fixed 1.24-V control voltage provided by v_{cont} eventually yields a 26- μm drawn channel length. Incidentally, it is advisable to also double check the outcome of each iteration on the closed-loop magnitude response after plotting with

```
plot vdb(4,3)-vdb(8)
```

Once the proper channel length has thus been determined, repeat the ac simulation after resetting the control voltage to 1.59 V as determined in design for shifting the corner frequency to 12 kHz. Surely, the corner frequency indicated by the simulation will be somewhat different from the target. Iterate by readjusting the control voltage until the 12-kHz target is achieved, and adopt the corresponding control voltage as true $V_{\text{cont(max)}}$. In the present example, it turns out to be 1.62 V.

The dc-analysis line appearing in the input file is for verifying the range performance of the filter. Plotting the derivative $\text{deriv}(v(4,3))$ of the differential output voltage displays the variation of the fully-differential dc gain with the input signal v_i . The outcome for $V_{\text{cont}} = 1.24 \text{ V}$ and $V_{\text{cont}} = 1.62 \text{ V}$ are shown in Figs. 3.52(c) and (d), respectively. Both plots indicate a reasonably constant 2-V/V dc gain over the specified input-signal range $-0.2(\text{V}) \leq v_i \leq 0.2(\text{V})$. The reader may quantify the nonlinearity within this range by running a transient simulation with a sinusoidal input signal of say 100-Hz frequency and reading out the total harmonic distortion. ▲

3.11 Capacitor Design 电容设计

As discussed in Sections 2.11 and 2.12, discrete-time amplifiers are implemented with capacitive external networks. Capacitors also find use as external-network components in continuous-time filtering applications, and as internal components in opamp stabilization by frequency compensation.

Advanced CMOS technologies offer a parallel-plate capacitor structure in the form of two metal layers separated by an insulator. Called MIM, this type of capacitor provides a highly linear capacitance. Another dedicated capacitor, constructed by sandwiching an insulator film in between two polysilicon layers, and called PIP, is offered in 0.25 μm and earlier CMOS technologies. However, the quality of PIP is inferior to MIM in linearity and capacitive/resistive parasitics. In those applications where nonlinearity can be tolerated, a MOSFET can be utilized as a capacitor of significantly larger capacitance per unit area than MIM or PIP. Decoupling is the most common of such applications. Stripped of its source and drain, a MOSFET can also be used as a voltage-controlled capacitor, that is, a varactor, which finds widespread use in frequency-tuning applications.

In this section, the modeling and ratioing aspects of MIM capacitors will be discussed first. Subsequently, the MOSFET will be explored as a decoupling capacitor and as a varactor.

3.11.1 MIM Capacitor MIM电容

The cross-sectional view of an MIM capacitor is given in Fig. 3.53(a), where metal layers T and $T - 1$ represent the two topmost standard metal layers available in CMOS technologies. The former of the two is reserved for accessing the upper and lower plates of the capacitor. The lower plate is built in metal $T - 1$, whereas a dedicated metal layer (cap metal) is used for building the upper plate. The plates are laid out typically in a square geometry with slightly shorter sides on the upper plate. The dimensions of the latter therefore defines the effective area and perimeter of the capacitor.

The capacitance of an MIM capacitor, represented by C in Fig. 3.53(a), is usually modeled by

$$C = (C_A A + C_F P) (1 + VCC \times |v_A - v_B|) [1 + TCC1 \times (T - T_0) + TCC2 \times (T - T_0)^2], \quad (3.137)$$

where A and P are respectively the area and perimeter of the upper plate, $|v_A - v_B|$ is the magnitude of the voltage across the capacitor, C_A and C_F are the coefficients of the

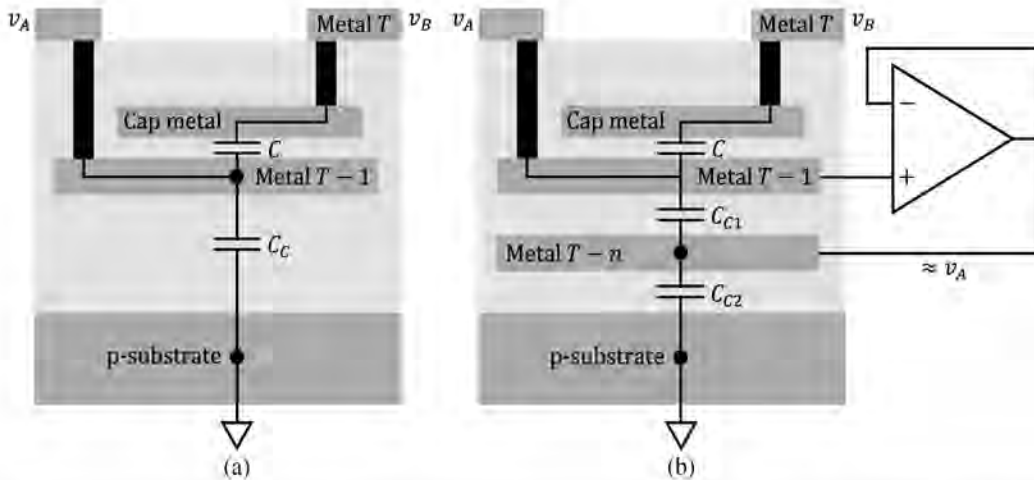


FIGURE 3.53 (a) Cross-sectional view of an MIM capacitor. Top metal layer (metal T) is used for accessing the two plates of the capacitor. Lower plate of the capacitor is built in Metal $T - 1$, which is the standard metal layer just below the top metal. Upper plate of the capacitor is built in a dedicated metal layer (cap metal). C denotes the main capacitance. C_c is the parasitic capacitance between the lower plate and substrate. (b) Screening with a lower metal layer (metal $T - n$) for minimizing the parasitic capacitance of the lower plate.

areal and fringe components of the capacitance, VCC is the voltage coefficient, $TCC1$ and $TCC2$ are temperature coefficients, T is the operation temperature, and T_0 is the temperature at which the capacitance parameters are characterized by the manufacturer. Numerical examples of MIM-capacitor parameters in a typical 0.18- μm CMOS technology are as follows: $C_A = 1 \text{ fF}/\mu\text{m}^2$, $C_F = 0.2 \text{ fF}/\mu\text{m}$, $VCC = 2 \times 10^{-5} \text{ V}^{-1}$, $TCC1 = 5 \times 10^{-3} \text{ C}^{-1}$, $TCC2 = 1 \times 10^{-5} \text{ C}^{-2}$, and $T_0 = 25^\circ\text{C}$.

The combination of inaccuracy and imprecision may be as large as 15% of the designed value of an MIM capacitance. The ratio of two MIM capacitances, however, exhibits negligible inaccuracy provided that common-centroid interdigitation and dummy termination are practiced in physical design. As for the imprecision of an MIM capacitance, the simplest model prescribes

$$\frac{\sigma[C]}{C} = \frac{K_C}{\sqrt{A}} \quad (3.138)$$

for the relative standard deviation, where A (μm^2) and K_C (μm) represent the area and Pelgrom coefficient, respectively. In 0.18- μm CMOS technologies, K_C is typically around 0.01 μm .

As shown in Fig. 3.53(a), the lower plate of an MIM capacitor is coupled to the substrate via a parasitic capacitance C_C , which can be as large as 5% of the main capacitance. This parasitic capacitance adversely affects the applications in which the lower plate is on the signal path. It may inject substrate noise onto the signal path or alter the closed-loop performance of *stray-sensitive* configurations. Fortunately, these effects can be suppressed by introducing a screening plate in between the bottom plate and the substrate. Any metal layer below $T-1$ can be used for this purpose. This arrangement is depicted in Fig. 3.53(b) with a metal layer labeled $T-n$. For noise suppression, it suffices to connect the screen to a clean ground node but this won't eliminate the effect on the closed-loop properties of a stray-sensitive configuration. As a matter of fact, the bottom plate will be more strongly coupled to ground because its parasitic capacitance C_{C1} to the screen is larger than the original parasitic capacitance C_C to the substrate. A radical technique by which the capacitive coupling of the bottom plate can be virtually eliminated is to drive the screen with the voltage v_A of the bottom plate. The OTA-based unity-gain amplifier shown in Fig. 3.53(b) accomplishes this task almost ideally. Considering the gain error of the amplifier, the voltage across C_{C1} is given by v_A/A_0 , where A_0 is the dc-differential gain of the OTA. The parasitic capacitance is therefore effectively reduced to C_{C1}/A_0 . Although the screen itself is capacitively coupled to the substrate, the capacitance C_{C2} associated with this coupling is driven by the buffer outside the signal path. This screening technique is of course expensive due to the power consumption and footprint of the amplifier. In less critical applications, the parasitic capacitance between the lower plate and substrate can be reduced simply by placing a floating n-well beneath the MIM capacitor. As described in Subsection 3.10.1 for minimizing the parasitic capacitance of poly resistors, this technique works by replacing C_C with $C_C C_W / (C_C + C_W)$ where C_W is the junction capacitance of the well.

Example 3.22 The external network of the fully-differential discrete-time amplifier of Fig. 2.52 is to be built with MIM capacitors of $C_A = 1 \text{ fF}/\mu\text{m}^2$, $C_F = 0.2 \text{ fF}/\mu\text{m}$, and

$K_C = 0.01 \mu\text{m}$. Determine the size and capacitance of all capacitors in accordance with the following specifications:

1. Target closed-loop gain: $A_{cl} = 7 \text{ V/V}$.
2. 4.5σ random external-network error: Equal to or smaller than 0.1% of the output signal.

Work Closed-loop gain A_{cl} is described by (2.212) as $1 + (C_1/C_2)$, where C_1 is the target value of C_{A1} and C_{B1} , and C_2 is the target value of C_{A2} and C_{B2} . Item 1 of specifications therefore implies

$$\frac{C_1}{C_2} = 6. \quad (3.139)$$

Considering (2.219), we can represent the 4.5σ random external-network error with $4.5\sigma[1/\beta']v_i(t_2)$, where β' is described by (2.216). On the other hand, the output signal is given by $A_{cl}v_i(t_2)$. Considering $A_{cl} = 7 \text{ V/V}$, item 2 of specifications can therefore be expressed as

$$4.5 \times \sigma [1/\beta'] \times v_i(t_2) \leq 0.001 \times 7 \times v_i(t_2),$$

hence

$$\sigma [1/\beta'] = 1.56 \times 10^{-3}. \quad (3.140)$$

The variance of $1/\beta'$ can be related to the variances of C_{A1} , C_{A2} , C_{B1} , and C_{B2} through (2.216), (2.28), and (3.139) as follows:

$$\sigma^2 [1/\beta'] = 9 \left(\frac{\sigma^2 [C_{A1}]}{C_{A1}^2} + \frac{\sigma^2 [C_{A2}]}{C_{A2}^2} + \frac{\sigma^2 [C_{B1}]}{C_{B1}^2} + \frac{\sigma^2 [C_{B2}]}{C_{B2}^2} \right). \quad (3.141)$$

Assuming that (3.139) is fulfilled by assigning one unit capacitor C_U to each of C_{A2} and C_{B2} and six parallel unit capacitors to each of C_{A1} and C_{B1} , the relative variances appearing on the right-hand side of (3.139) can be expressed as

$$\begin{aligned} \frac{\sigma^2 [C_{A1}]}{C_{A1}^2} &= \frac{\sigma^2 [C_{B1}]}{C_{B1}^2} = \frac{1}{6} \frac{\sigma^2 [C_U]}{C_U^2}, \\ \frac{\sigma^2 [C_{A2}]}{C_{A2}^2} &= \frac{\sigma^2 [C_{B2}]}{C_{B2}^2} = \frac{\sigma^2 [C_U]}{C_U^2}, \end{aligned}$$

which turn (3.141) into

$$\sigma^2 [1/\beta'] = 21 \frac{\sigma^2 [C_U]}{C_U^2}.$$

Combining this equation with (3.138) and (3.140) and adopting $K_C = 0.01 \mu\text{m}$ as specified, we finally obtain $A = 864 \mu\text{m}^2$ for the area of a unit capacitor. For a square capacitor layout, this leads to a perimeter $P = 118 \mu\text{m}$. The nominal value of unit capacitance, as calculated from (3.137) for zero voltage and $T = T_0$, is therefore $C_U = 888 \text{ fF}$. ▲

3.11.2 MOSFET as a Capacitor MOS晶体管用作电容

A capacitor built out of a MOSFET is called MOSCAP. The capacitance of a MOSCAP is typically an order-of-magnitude higher than that of a MIM capacitor of comparable size. However, it is hardly linear, which is why its utility is mostly in decoupling applications where linearity is secondary to the density of capacitance per unit area.

A MOSFET is configured into a MOSCAP by combining its source and drain into a common terminal called *source*. The consequence $v_{DS} = 0$ of this combination turns the general high-frequency small-signal equivalent circuit of Fig. 3.41 into the three-terminal form depicted in Fig. 3.54(a). Another implication of $v_{DS} = 0$ is $C_{gb} = C_{bg}$, which further transforms the equivalent circuit into what is shown in Fig. 3.54(b). Note that this final form contains three physical interterminal capacitances defined by $C_{GB} \doteq C_{bg}$, $C_{GS} \doteq C_{gg} - C_{bg}$, and $C_{SB} \doteq C_{bb} - C_{bg}$. Recalling (3.112) for C_{bg} , (3.105) for C_{gg} , and (3.111) for C_{bb} , and subsequently using (3.99) for C_{ggs} , (3.76) for C_{ggs} , (3.101) for C_{bbe} , and (3.78) for C_{bbi} , we arrive at the following expressions for the capacitances of a MOSCAP:

$$C_{GB} = C_{gbe} - C_{bgi}, \quad (3.142)$$

$$C_{GS} = C_{gse} + C_{gde} - (C_{gsi} + C_{gdi}), \quad (3.143)$$

$$C_{SB} = C_{sbe} + C_{dbe} - (C_{bsi} + C_{bdi}). \quad (3.144)$$

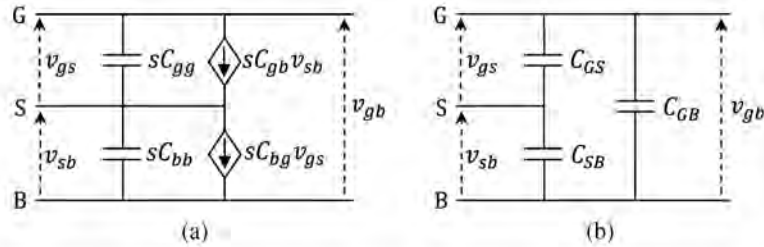


FIGURE 3.54 (a) MOSCAP equivalent circuit as derived from Fig. 3.41 for $v_{ds} = 0$. (b) MOSCAP equivalent circuit represented by physical interterminal capacitances only.

Extrinsic components are related to device dimensions through (3.91) to (3.95). Intrinsic components, on the other hand, are proportional to gate area with an areal density described by (3.89), and are strongly dependent on terminal voltages as observed in Figs. 3.55(a) to (c) on simulated density plots. For an explanation of the physical roots of this strong nonlinearity, consider Fig. 3.55(d) first. It shows a MOSCAP channel operating in subthreshold and therefore supporting only a depletion layer at the silicon surface as highlighted in light gray. This layer, having much lower carrier concentrations in comparison with the source and substrate regions, acts as a dielectric and forms a *depletion capacitance* C_{st} between silicon surface and bulk terminal. The gate-oxide layer overlying the channel is also a dielectric. It forms an *oxide capacitance* C_{oxide} in between gate and silicon surface. These two capacitances quite generally account for all the intrinsic capacitances appearing in (3.142) to (3.144). However, due to the diminishing electron population of the channel in subthreshold, none of these two capacitances is connected to the n-type source terminal. This is why the intrinsic components plotted in Figs. 3.55(a) and (b) for the source-related capacitances C_{GS} and C_{SB} are vanishingly small under the subthreshold condition $V_{GB} < V_{SB} + V_{TN}$. In contrast, Fig. 3.55(c) indicates a maximized intrinsic component for C_{GB} under the same condition. This is because the depletion and

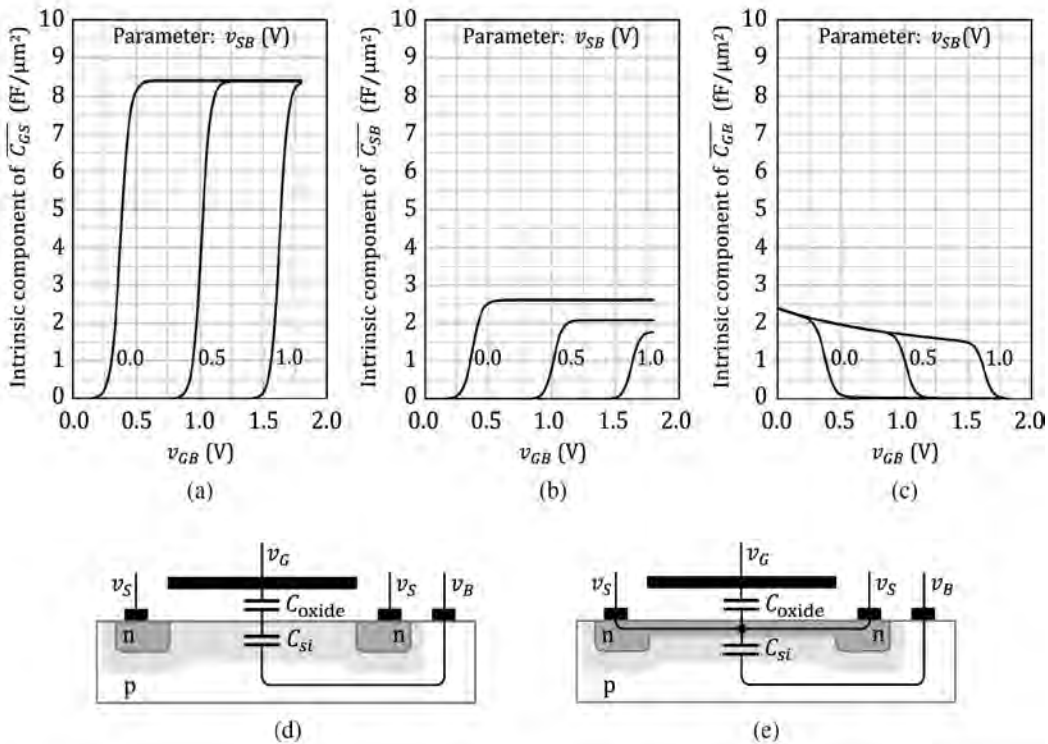


FIGURE 3.55 (a) Gate-source intrinsic capacitance $-(\overline{C_{gst}} + \overline{C_{gdl}})$. (b) Source-bulk intrinsic capacitance $-(\overline{C_{bsl}} + \overline{C_{bdl}})$. (c) Gate-bulk intrinsic capacitance $-\overline{C_{bgl}}$. (d) Network of capacitances in subthreshold. (e) Network of capacitances in strong inversion.

oxide capacitances are acting in series between the gate and bulk terminals as depicted in Fig. 3.55(d).

Now consider Fig. 3.55(e), where a strongly inverted channel is depicted. The dark gray sheet at the surface of the channel represents the inversion layer of electrons, which connects the common node of the oxide and depletion capacitances to the source terminal. Due to this connection, the oxide capacitance now appears in between gate and source terminals, and thus maximizes the intrinsic component of C_{GS} as observed in Fig. 3.55(a) for $V_{GB} > V_{SB} + V_{TN}$. Since the gate terminal is no more coupled to the bulk, the intrinsic component of C_{GB} vanishes in strong inversion as observed in Fig. 3.55(c). Finally, notice in Fig. 3.55(b) the maximized intrinsic component of C_{SB} . This is due to the fact that the depletion capacitance is now in between the source and bulk terminals.

In decoupling applications, the node to be decoupled is tied to the ground via a large capacitance. The most popular MOSCAP configuration serving this purpose is shown in Fig. 3.56(a). Notice the grounding of source and bulk terminals. Gate is tied to the node to be decoupled. The effective capacitance C_G between gate and ground equals $C_{GS} + C_{GB}$. As illustrated in Fig. 3.56(b), the density of the intrinsic component of this capacitance is maximized in strong inversion to the density C_{OX} of the oxide capacitance. The decoupled voltage v_G should therefore be kept above the threshold voltage in order to maintain this maximum value.

A capacitor serving as a voltage-controlled capacitance is known as *varactor*. The most popular varactor structure is shown in Fig. 3.57(a). Notice the absence of a source terminal. Gate and n-type bulk are the two terminals of its capacitance C_{GB} , which equals the series equivalent of the gate-oxide capacitance C_{oxide} and the depletion capacitance

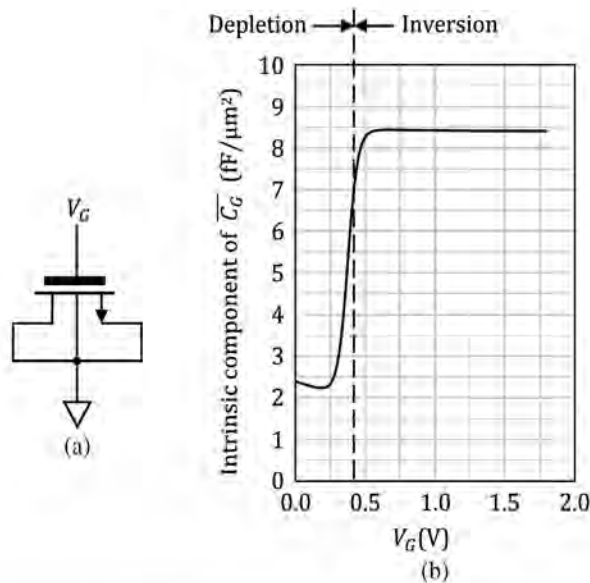


FIGURE 3.56 (a) A decoupling MOSCAP (b) Areal density of the intrinsic components of the capacitance between gate and ground.

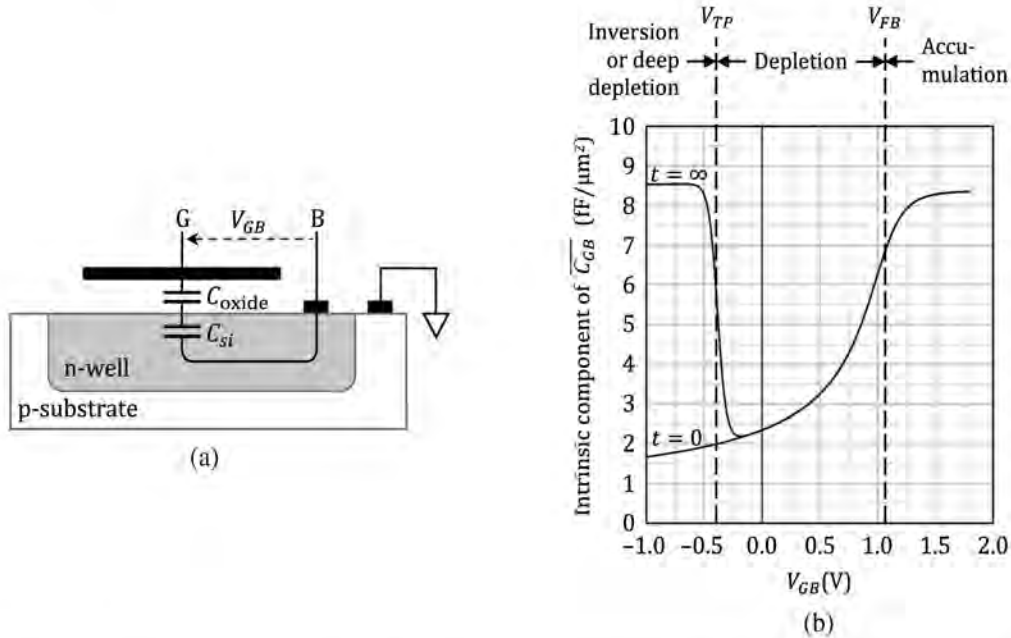


FIGURE 3.57 (a) Varactor structure. (b) The variation of the areal density of gate-bulk capacitance with gate-bulk voltage.

C_{si} , C_{GB} is controlled by the gate-bulk voltage v_{GB} as shown in Fig. 3.57(b) by a simulated plot of its areal density $\overline{C_{GB}}$.

For a description of varactor operation, consider the case $V_{GB} < V_{TP}$, for which, a regular p-channel MOSCAP built inside this n-well with a source-bulk short would be in strong inversion. In a varactor, however, inversion is conditional. When V_{GB} is lowered below V_{TP} rapidly, the surface of the n-well initially goes into a *deep depletion* state instead of being inverted because, unlike a MOSCAP, a varactor does not have a p-type source terminal from which the holes needed for inversion can be supplied immediately. Holes are supplied to the surface at a very slow pace after being thermally generated inside the n-type bulk. The replacement of the deeply depleted surface layer by an inversion layer therefore takes time. The initially large width of the depletion layer gradually shrinks in a transient state and eventually disappears throughout this process. As a result, depletion capacitance evolves over time from a very small value toward infinity. This is why $\overline{C_{GB}}$ also begins with a small initial value labeled $t = 0$ in Fig. 3.57(b), and increases toward C_{OX} as labeled $t = \infty$. Since this delay precludes the instantaneous control of C_{GB} , it must be avoided by not letting v_{GB} drop below V_{TP} . Raising v_{GB} above V_{TP} brings the surface into a regular depletion mode, in which, response to v_{GB} is not delayed because the surface is no more in need of the scarcely available holes. The width of this depletion region is a decreasing function of v_{GB} , which is why the depletion capacitance and therefore C_{GB} increase with v_{GB} . This trend comes to an end as the depletion layer vanishes for $v_{GB} = V_{FB}$, where V_{FB} is called *flatband voltage*. For $v_{GB} > V_{FB}$, an *accumulation* layer of electrons prevails at the surface. These electrons effectively connect the lower node of

the oxide capacitance to the bulk terminal, and thus maximize $\overline{C_{GB}}$ at C_{OX} . Since n-well can instantaneously supply the electrons needed by the accumulation layer, response to v_{GB} in this mode is also free of delay.

It is obvious from the foregoing description that the range of v_{GB} , in which, C_{GB} can be varied is limited approximately by V_{TP} at the low end and by V_{FB} at the high end. As observed in Fig. 3.57(b), this range is no less than a volt wide in a typical 0.18- μm CMOS technology. Incidentally, if V_{FB} is not quantified in a BSIM parameter deck, it can still be estimated from

$$V_{FB} \approx V_{TH0} + 2\phi_{FB} + K1 \sqrt{2\phi_{FB}}. \quad (3.145)$$

3.12 References 参考文献

1. Y. S. Chauhan, S. Venugopalan, M. A. Karim, S. Khandelwal, N. Paydavosi, P. Thakur, A. M. Niknejad, et al., "BSIM—Industry standard compact MOSFET models," in *Proceedings of European Solid-State Device Research Conference*, pp. 46–49, 2012.
2. BSIM Group. bsim.berkeley.edu/ (accessed November 1, 2017).
3. The MOSIS Service. www.mosis.com/Technical/Testdata/tsmc-018-prm.html (accessed March 15, 2010).
4. J. Shin, I.-Y. Chung, Y. J. Park, and H. S. Min, "A new charge pump without degradation in threshold voltage due to body effect," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1227–1230, 2000.
5. BSIM Group. <http://bsim.berkeley.edu/models/bsim3/> (accessed February 12, 2018).
6. BSIM Group. <http://bsim.berkeley.edu/models/bsim4/> (accessed February 12, 2018).
7. BSIM Group. <http://bsim.berkeley.edu/models/bsimbulk/> (accessed February 12, 2018).
8. R. Thewes, R. Brederlow, C. Dahl, U. Kollmer, C. G. Linnenbank, B. Holzapfl, J. Becker, et al., "Explanation and quantitative model for the matching behaviour of poly-silicon resistors," *Technical Digest, International Electron Devices Meeting*, pp. 771–774, 1998.

3.13 Exercises 习题

3.1 A model deck for standard-threshold devices of 90-nm bulk CMOS technology is provided in `bsim90.sp`. It is a slightly modified version of a model deck prepared by the Nanoscale Integration and Modeling (NIMO) Group at Arizona State University and made public at the following URL of Predictive Technology Model:

http://ptm.asu.edu/modelcard/90nm_bulk.txt

For this technology, assume $L_{DR(\min)} = 0.1 \mu\text{m}$ and $W_{DR(\min)} = 0.3 \mu\text{m}$ for minimum permissible dimensions, and one half of the numerical values cited in Table 3.2 for physical design rules. Also assuming a maximum voltage of 1.2 V instead of 1.8 V, extract the design tools appropriate for this technology. When rewriting `nmos1.sp`, `nmos2.sp`, `pmos1.sp`, and `pmos2.sp` for this purpose, it is appropriate to scale down the dimensional variables by a factor of 2. Scaling down the port voltages by a factor around 1.5 is also meaningful since their upper limit is reduced from 1.8V to 1.2V. Make a record of the extracted design tools because they will be utilized in some of Chapter 4 exercises. It is worth mentioning that the threshold-voltage tools will not exhibit a monotonic relationship between threshold voltages and channel length. As observed in some short-channel bulk CMOS technologies, V_{TN} and $|V_{TP}|$ exhibit a maximum for a channel length close to $L_{DR(\min)}$.

3.2 An NMOS transistor is to operate as a current sink with $V_{SB} = 0$. Its performance is specified by $V_{DSsat} = 200 \text{ mV}$, $I_D = 100 \mu\text{A}$, and $r_{ds} = 200 \text{ k}\Omega$ at $V_{DS} = 0.9 \text{ V}$. Set up and apply a design plan for determining V_{GS} , L_{DR} , and W_{DR} .

3.3 A PMOS transistor is to operate as a transconductor with $V_{SB} = -300 \text{ mV}$. Its performance is specified by $V_{DSsat} = -200 \text{ mV}$, $g_m = 600 \mu\text{A/V}$, and $r_{ds} = 200 \text{ k}\Omega$ at $V_{DS} = -0.9 \text{ V}$. Set up and apply a design plan for determining V_{GS} , L_{DR} , and W_{DR} .

3.4 A MOSFET is specified by V_{GS} , V_{SB} , V_{DS} , I_D , and r_{ds} . Which steps do you take for determining L_{DR} and W_{DR} ?

3.5 The PMOS transistor shown in Fig. 3.58 is to operate in saturation with a source-bulk short. Determine V_B as well as W_{DR} , L_{DR} in accordance with the following specifications: $I = 150 \mu\text{A}$, $g_m = 1.5 \text{ mA/V}$, $r_{ds} = 100 \text{ k}\Omega$ for $V_O = 0.9 \text{ V}$.

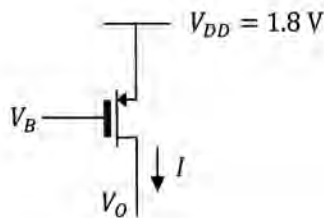


FIGURE 3.58

3.6 The NMOS transistors shown in Fig. 3.59 have identical channel lengths and identical channel widths. They are biased by dc gate voltages V_{B1} and V_{B2} , and conduct a dc current $I = 1 \text{ mA}$ in saturation. While M1 operates with a source-bulk short, M2 doesn't. The drain-source saturation voltage of M1 happens to be 125 mV.

Determine (a) the drain-source saturation voltage of M2, (b) overdrive voltage of M1, (c) overdrive voltage of M2, (d) sheet current of M1, (e) sheet current of M2, (f) aspect ratio of M1, and (g) aspect ratio of M2. Also derive by inspection the relationship between dc small-signal output resistance $r_o \doteq v_o/i$ and dc small-signal device parameters.

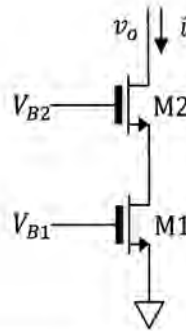


FIGURE 3.59

3.7 V_{B1} and V_{DD} shown in Fig. 3.60 are dc voltage sources. All devices operate with source-bulk shorts. Derive by inspection the relationship between dc small-signal voltage gain v_o/v_i and dc small-signal device parameters.

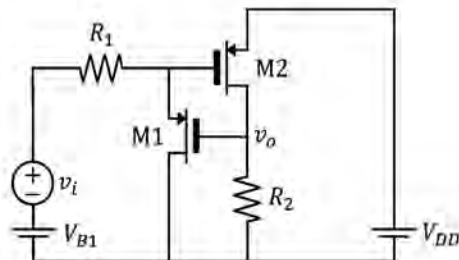


FIGURE 3.60

3.8 Shown in Fig. 3.61 is the circuit schematic of a source-follower stage. The MOSFET operates without a source-bulk short. It is specified by small-signal parameters g_m , g_{mb} , and r_{ds} . V_{DD} , V_{SS} , and V_I are supplied by dc voltage sources. I_B is a dc current source.

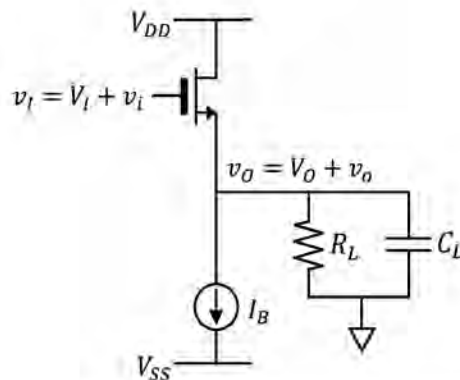


FIGURE 3.61

$v_I = V_I + v_i$ is the input voltage, where V_I is the dc bias, and v_i is the signal. Output voltage is represented by $v_O = V_O + v_o$, where V_O is the bias component, and v_o is the signal component. Determine by inspection the dc small-signal voltage gain v_o/v_i in terms of device parameters.

3.9 Shown in Fig. 3.62 is the circuit schematic of a differential amplifier stage. All MOSFETs operate *with* source-bulk shorts. M1 and M2 are identical devices specified by small-signal parameters g_m , g_{mb} , and r_{ds} . M3 and M4 are also identical. They are specified by g'_m , g'_{mb} , and r'_{ds} . V_{DD} and V_{SS} are supplied by dc voltage sources. I_B is a dc current source. v_{i1} and v_{i2} are two voltage signal sources. v_o is the output signal. Construct by inspection a parametric expression for the dc small-signal v_o as a function of v_{i1} and v_{i2} in terms of device parameters.

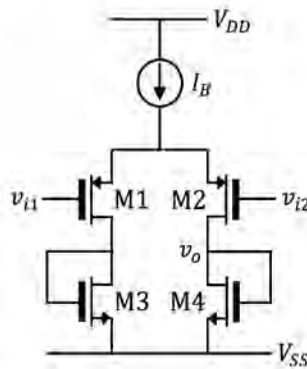


FIGURE 3.62

3.10 Shown in Fig. 3.63 is the circuit schematic of an amplifier stage with a feedback loop. All MOSFETs operate with source-bulk shorts. Small-signal parameters are specified as follows: g_{m1} , g_{mb1} , and r_{ds1} for M1; g_{m2} , g_{mb2} , and r_{ds2} for M2; g_{m3} , g_{mb3} , and r_{ds3} for M3. V_{DD} and V_{SS} are supplied by dc voltage sources. v_i and v_o represent input and output signals, respectively. Determine by inspection the dc small-signal voltage gain v_o/v_i in terms of device parameters.

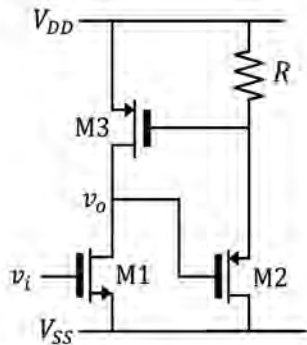


FIGURE 3.63

3.11 An amplifier stage used occasionally as a buffer is shown in Fig. 3.64. V_{SS} and V_{DD} are dc power-supply voltages, V_I is supplied by a dc bias source, and V_O is the dc bias voltage at the output. Prove that the following dc small-signal voltage-gain expression is valid irrespective of device dimensions:

$$\frac{v_o}{v_i} = \frac{V_{GS2} - V_{TN}}{V_{GS1} - V_{TN}}$$

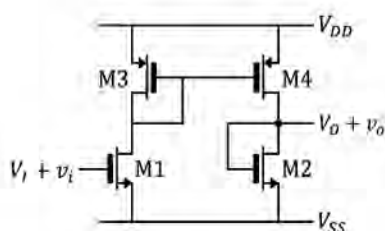


FIGURE 3.64

3.12 A gain stage used in class-AB opamps is shown in Fig. 3.65. It contains a driver device M1, a bias device M4, and two diode-connected devices M2 and M3. Derive by inspection an expression for the dc small-signal voltage gain v_o/v_i in terms of device transconductances and output resistances. Simplify the expression by assuming any g_m to be much larger than any $1/r_{ds}$.

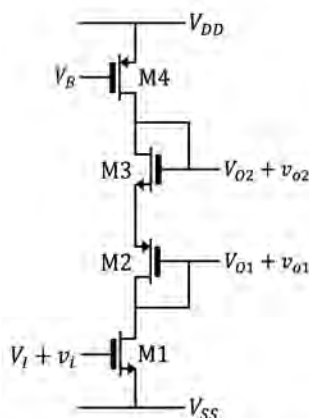


FIGURE 3.65

3.13 Shown in Fig. 3.66 is the circuit schematic of an amplifier stage built with a self-biased inverter. Both MOSFETs operate with source-bulk shorts. Draw the high-frequency small-signal equivalent circuit of this stage without showing ineffective components.

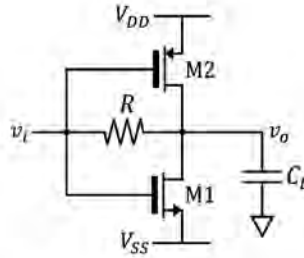


FIGURE 3.66

3.14 Shown in Fig. 3.67 are two ratioed high-resistivity poly resistors that are segmented into matching unit resistors and laid out with a common centroid. Contacts are represented with black squares, and metal interconnections are drawn in gray. Mark the two terminals of one of the resistors with R_A and those of the other resistor with R_B . Also place a cross where you believe the common centroid is located.

1. Determine the target resistance of resistor R_A by assuming $R_C = 10\ \Omega$ for contact resistance and referring to Table 3.4 for other resistance parameters.
2. Repeat part 1 for R_B .
3. Determine R_A/R_B .
4. Assuming $K_R = 0.04\ \mu\text{m}$ for the Pelgrom coefficient, evaluate the relative standard deviation

$$\frac{\sigma[R_A/R_B]}{R_A/R_B}$$

of the resistance ratio.

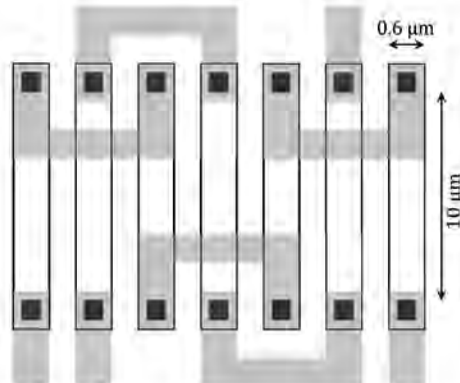


FIGURE 3.67

3.15 As explained in Subsection 3.11.1 and modeled by (3.137), the capacitance of an MIM capacitor has an areal component proportional to A and a fringe component proportional to P . If you wish to accurately establish a certain capacitance ratio $C_2/C_1 = S$ between two MIM capacitors without resorting to segmentation, you need to make sure that both A_2 and P_2 are scaled with respect to A_1 and S_1 by the same S . Now suppose C_1 has a square layout of 10- μm side length, and determine the dimensions of a rectangular C_2 for $S = 1.26$.

