

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
+CGS0 = 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 CMOSPMOS (
+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
+CGS0 = 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
A0 = 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
CGB0 = 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
CGD0 = 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.

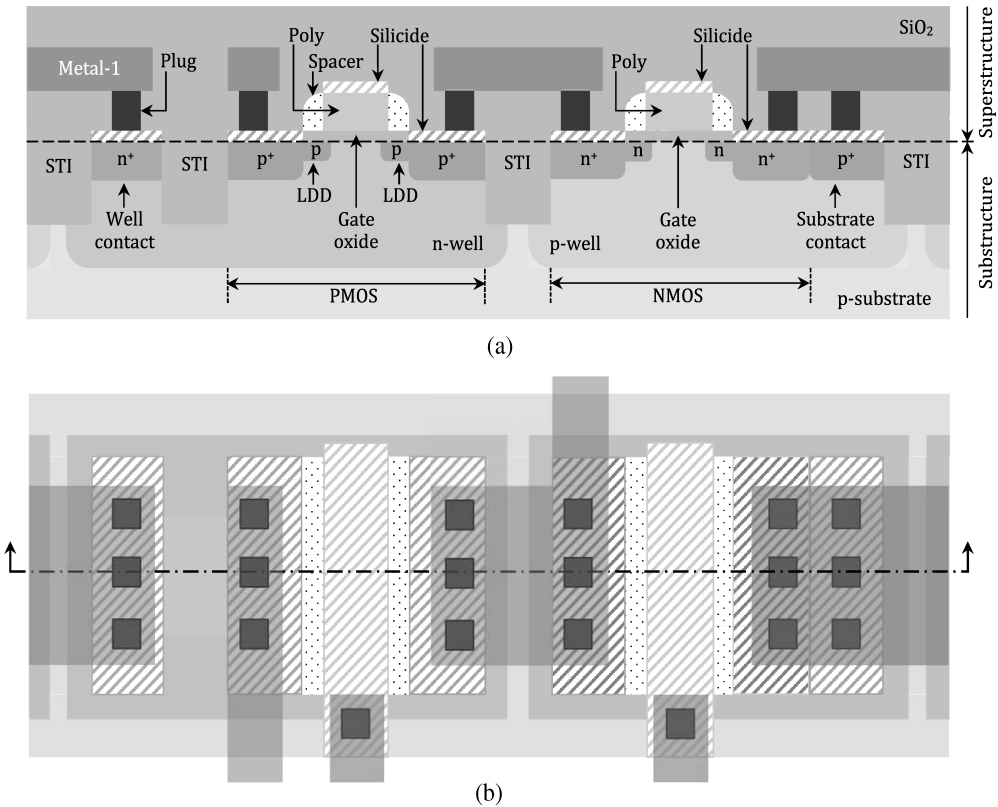


FIGURE 3.1 Twin-well bulk CMOS structure. (a) Cross-sectional view. (b) Top view and line of sectioning.

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

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

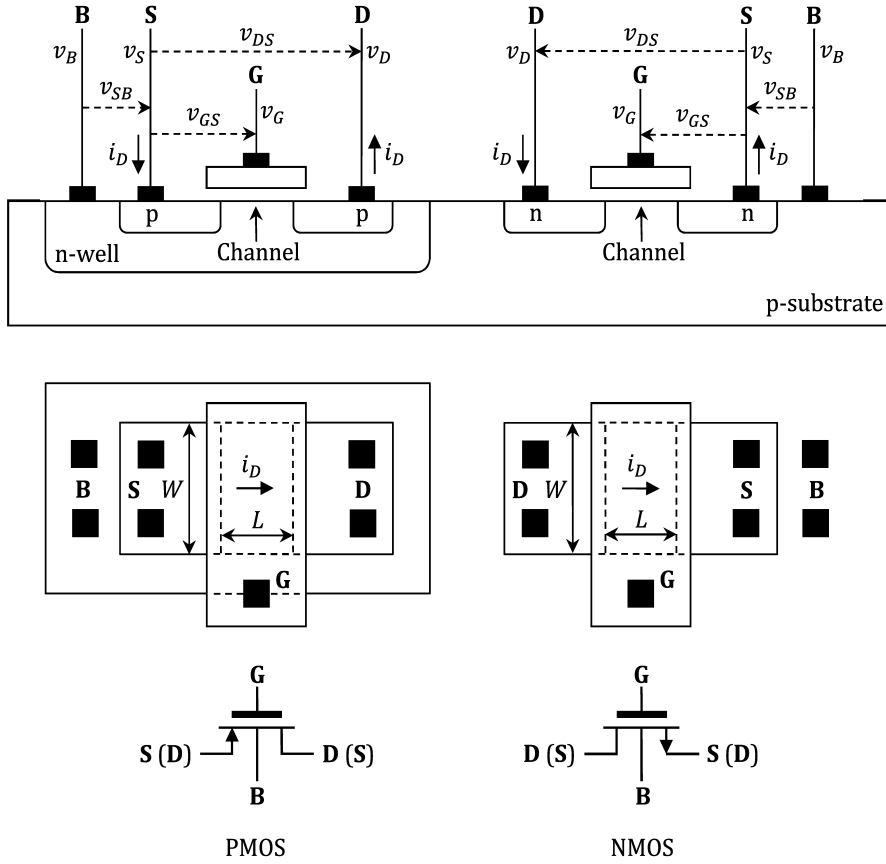


FIGURE 3.2 Top: Simplified cross-sectional view and electrical port variables of NMOS and PMOS structures. Middle: Top view of the structures. Bottom: Symbols representing these devices in circuit schematics.

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 $v_{SB} \doteq v_S - v_B$. 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.

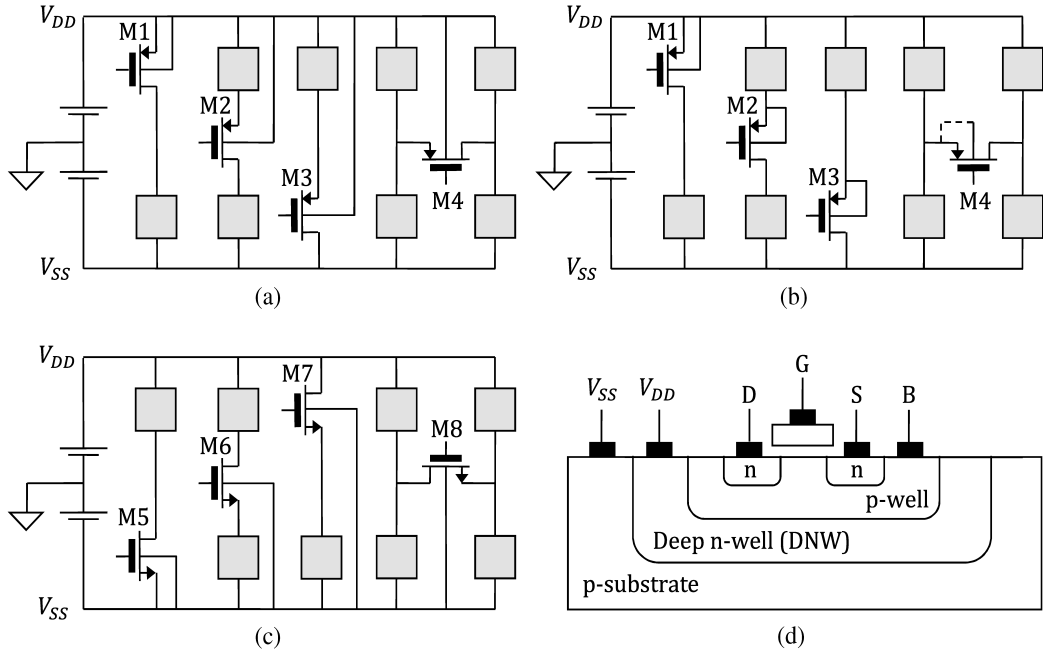


FIGURE 3.3 (a) All PMOS devices sharing a common bulk at V_{DD} . (b) Possibility of individual bulk-source connection in PMOS devices M1, M2, and M3. M4 can have an individual bulk-source connection if source designation does not change during operation. Otherwise, bulk is tied to V_{DD} . (c) All NMOS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

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