ENGR 2420: Introduction to Microelectronic Circuits

March 10, 2020

4 MOS Transistor Models

In this section, we shall develop large-signal models for an nMOS transistor in all of its various operating regimes. Because the MOS transistor is a four-terminal device, the current depends on three potential differences, which are typically taken to be V_{GS}, V_{DS}, and V_{BS}. Here, the gate, drain, and bulk (i.e., substrate) voltages are all referenced to that of the source. Although this convention is ubiquitous, it is not the only possible convention. For example, we can instead reference the gate, source, and drain potentials instead to the bulk potential. In this case, the model would be expressed in terms of V_{GB}, V_{SB}, and V_{DB} [1]. With this convention, the channel current of an MOS transistor can be expressed as

$$I = I_{\mathrm{s}} \left(\mathrm{f}(V_{\mathrm{GB}}, V_{\mathrm{SB}}) - \mathrm{f}(V_{\mathrm{GB}}, V_{\mathrm{DB}}) \right), \tag{1}$$

where I_{s} is related to the channel current of the transistor at threshold and $\mathrm{f}(\cdot)$ is a function that assumes an exponential form below threshold and a quadratic form above threshold. Such MOS models are called *bulk-referenced* for obvious reasons or *source/drain-symmetric* because the source and drain voltages come into the model in a symmetric fashion. With the proper choice of $\mathrm{f}(\cdot)$, this single equation describes the channel current in *all* regions of operation, transitioning smoothly from weak inversion to strong inversion and from the ohmic region to the saturation region. For an nMOS transistor fabricated in an n-well technology, as shown in Fig. 1, the p-type bulk is connected to ground, so the three potentials in the model would simply be V_{G}, V_{D}, and V_{S}.

Our intent in developing these models will not be physical rigor. Instead, our concern is to develop *simple* large-signal models that are useful from a circuit-design point of view, as opposed to a device-physics one. We will define some notions associated with MOS transistor operation, such as threshold and the onset of saturation in slightly unconventional ways so that we can develop the models with a minimum of physical detail and so that the concepts can be applied consistently in all operating regions. In our discussion, we will make several simplifying assumptions. First, we shall be assuming a long-channel device. Second, we shall assume that the device behavior is entirely uniform along the direction of its width (i.e., in the y direction) and that the mobile charge is located at the surface of the device, so that we can consider current flow in one spatial dimension, along the length of the channel (i.e., in the x direction). Third, we shall assume that there is no charge trapped in surface states. Fourth, we shall neglect contact potentials that exist between different materials.

4.1 Channel Capacitance

We shall take the conduction band edge deep in the silicon substrate to be the zero of potential. When the gate voltage is equal to the flatband potential, V_{fb}, then the conduction and valence bands are flat from deep within the silicon substrate all the way to the surface and there is no depletion region beneath the gate. Likewise, the bands are flat within the

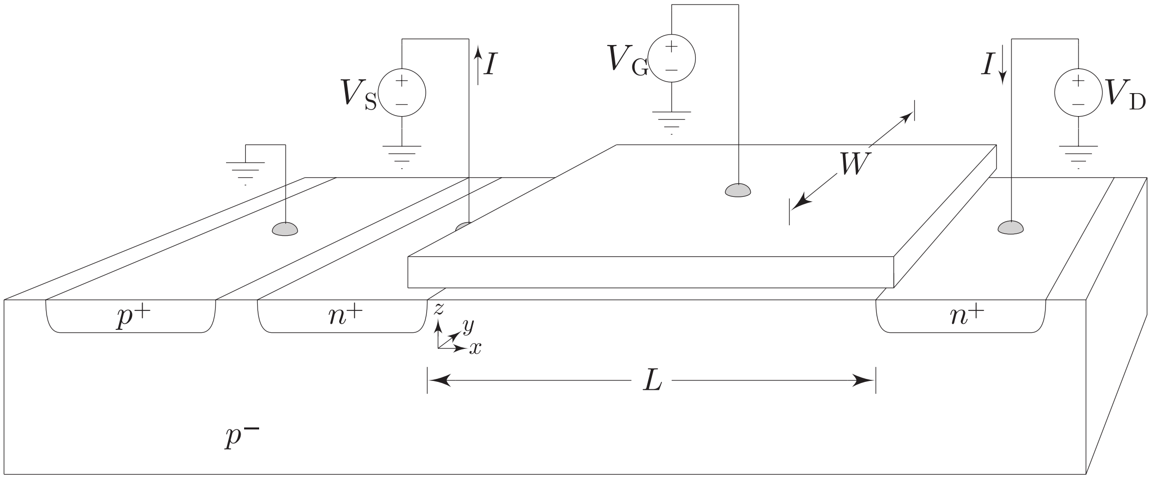


Figure 1: Simplified structure of an nMOS transistor.

oxide. If we apply a voltage to the gate that is more positive than V_fb, then the difference between the applied gate voltage and V_fb is partially dropped across the oxide and partially across a depletion region that forms beneath the gate, such that

$$V_\mathrm{G} - V_\mathrm{fb} = \psi_\mathrm{s} + \psi_\mathrm{ox}, \tag{2}$$

where ψ_s is the potential drop across the depletion region beneath the gate and ψ_ox is the potential drop in the oxide.

If there is no charge within the oxide, then Gauss' law implies that the electric field within the oxide, $\mathcal{E}_\mathrm{ox}$, is constant and is given by

$$\mathcal{E}_\mathrm{ox} = -\frac{\psi_\mathrm{ox}}{t_\mathrm{ox}}, \tag{3}$$

where t_ox is the thickness of the oxide layer. Integrating the charge density in the depletion layer and the inversion layer from deep within the silicon to the surface, we find the electric field at the surface of the silicon, $\mathcal{E}_\mathrm{si}$, to be

$$\mathcal{E}_\mathrm{si} = \frac{Q_\mathrm{dep} + Q_\mathrm{m}}{\epsilon_\mathrm{si}}, \tag{4}$$

where Q_dep is the charge per unit area in the depletion layer beneath the gate, Q_m is the mobile charge per unit area in the channel, and ϵ_si is the dielectric constant of silicon. The displacement vector (i.e., the product of the dielectric constant and electric field) is conserved across the silicon/oxide interface, which implies that

$$\epsilon_\mathrm{si}\mathcal{E}_\mathrm{si} = \epsilon_\mathrm{ox}\mathcal{E}_\mathrm{ox}.$$

Substituting Eqs. 3 and 4 into this equation and solving for ψ_ox, we find that

$$\psi_\mathrm{ox} = -\frac{t_\mathrm{ox}}{\epsilon_\mathrm{ox}}\left(Q_\mathrm{dep} + Q_\mathrm{m}\right) = -\frac{Q_\mathrm{dep} + Q_\mathrm{m}}{C_\mathrm{ox}}, \tag{5}$$

where $C_{\mathrm{ox}} \equiv \epsilon_{\mathrm{ox}}/t_{\mathrm{ox}}$ is the capacitance per unit area of the oxide. Substituting this equation into Eq. 2, we find that

$$V_{\mathrm{G}} - V_{\mathrm{fb}} = \psi_{\mathrm{s}} - \frac{Q_{\mathrm{dep}} + Q_{\mathrm{m}}}{C_{\mathrm{ox}}}. \tag{6}$$

Differentiating this equation with respect to distance along the channel, we find that

$$\begin{aligned} 0 &= \frac{\partial \psi_{\mathrm{s}}}{\partial x} - \frac{1}{C_{\mathrm{ox}}} \cdot \frac{\partial Q_{\mathrm{dep}}}{\partial \psi_{\mathrm{s}}} \cdot \frac{\partial \psi_{\mathrm{s}}}{\partial x} - \frac{1}{C_{\mathrm{ox}}} \cdot \frac{\partial Q_{\mathrm{m}}}{\partial x} \\ &= \frac{\partial \psi_{\mathrm{s}}}{\partial x} + \frac{C_{\mathrm{dep}}}{C_{\mathrm{ox}}} \cdot \frac{\partial \psi_{\mathrm{s}}}{\partial x} - \frac{1}{C_{\mathrm{ox}}} \cdot \frac{\partial Q_{\mathrm{m}}}{\partial x}, \end{aligned}$$

where $C_{\mathrm{dep}} \equiv -\partial Q_{\mathrm{dep}}/\partial \psi_{\mathrm{s}}$ is the incremental capacitance per unit area of the depletion layer beneath the gate. By rearranging this equation, we find that

$$\frac{\partial Q_{\mathrm{m}}}{\partial x} = \left(C_{\mathrm{ox}} + C_{\mathrm{dep}}\right) \frac{\partial \psi_{\mathrm{s}}}{\partial x} = C \frac{\partial \psi_{\mathrm{s}}}{\partial x}, \tag{7}$$

where $C \equiv C_{\mathrm{ox}} + C_{\mathrm{dep}} = \partial Q_{\mathrm{m}}/\partial \psi_{\mathrm{s}}$ is the incremental capacitance per unit area of the channel.

4.2 Channel Current

The channel current in an MOS transistor flows both by drift and by diffusion. In some operating regimes, the current flows primarily by drift, in others it flows primarily by diffusion. However, in many cases of interest, both current components are of roughly equal magnitudes. The predominant carrier transport mechanism can also change as a function of position along the channel in some regions of operation. Thus, to model the channel current in all regimes, we need to include both a drift and a diffusion term in the current-flow equation. Thus, we write the channel current, I, as a function of position along the channel, as

$$I(x) = I_{\mathrm{drift}}(x) + I_{\mathrm{diff}}(x). \tag{8}$$

The drift component of I is given by

$$I_{\mathrm{drift}}(x) = \left(W Q_{\mathrm{m}}\right) \mu \mathcal{E}_x = -W \mu Q_{\mathrm{m}} \frac{\partial \psi_{\mathrm{s}}}{\partial x}, \tag{9}$$

where W is the width of the MOS transistor (note that $W Q_{\mathrm{m}}$ gives the mobile charge density per unit length along the channel), μ is the effective low-field mobility of electrons in the channel, and $\mathcal{E}_x$ is the component of the electric field in the x direction (i.e., parallel to the channel). The diffusion component of I is given by

$$I_{\mathrm{diff}}(x) = D \frac{\partial}{\partial x} \left(W Q_{\mathrm{m}}\right) = W \mu U_{\mathrm{T}} \frac{\partial Q_{\mathrm{m}}}{\partial x}, \tag{10}$$

where D is the diffusion constant of electrons in the channel, and U_{T} is the thermal voltage, kT/q. Thus, we can rewrite Eq. 8 using Eqs. 9 and 10 as

$$\begin{aligned} I(x) &= -W \mu Q_{\mathrm{m}} \frac{\partial \psi_{\mathrm{s}}}{\partial x} + W \mu U_{\mathrm{T}} \frac{\partial Q_{\mathrm{m}}}{\partial x} \\ &= W \mu \left(-Q_{\mathrm{m}} \frac{\partial \psi_{\mathrm{s}}}{\partial x} + U_{\mathrm{T}} \frac{\partial Q_{\mathrm{m}}}{\partial x}\right). \end{aligned} \tag{11}$$

To obtain an expression for the channel current that is independent of position along the channel, we integrate both sides of Eq. 11 from source to drain. Doing so, we write

$$\int_0^L I(x)\,dx = W\mu \int_0^L \left(-Q_{\mathrm{m}}\frac{\partial\psi_{\mathrm{s}}}{\partial x} + U_{\mathrm{T}}\frac{\partial Q_{\mathrm{m}}}{\partial x}\right) dx.$$

Now, conservation of charge implies that the total channel current must be constant along the channel, so the left-hand side of this equation becomes

$$\int_0^L I(x)\,dx = I\int_0^L dx = IL.$$

Thus, we can write the channel current as

$$\begin{aligned}
I &= \frac{W}{L}\mu \int_0^L \left(-Q_{\mathrm{m}}\frac{\partial\psi_{\mathrm{s}}}{\partial x} + U_{\mathrm{T}}\frac{\partial Q_{\mathrm{m}}}{\partial x}\right) dx \\
&= S\mu \int_0^L \left(-Q_{\mathrm{m}}\frac{\partial\psi_{\mathrm{s}}}{\partial Q_{\mathrm{m}}}\cdot\frac{\partial Q_{\mathrm{m}}}{\partial x} + U_{\mathrm{T}}\frac{\partial Q_{\mathrm{m}}}{\partial x}\right) dx \\
&= S\mu \int_0^L \left(-\frac{Q_{\mathrm{m}}}{C} + U_{\mathrm{T}}\right)\frac{\partial Q_{\mathrm{m}}}{\partial x} dx \\
&= S\mu \int_{Q_{\mathrm{S}}}^{Q_{\mathrm{D}}} \left(-\frac{Q_{\mathrm{m}}}{C} + U_{\mathrm{T}}\right) dQ_{\mathrm{m}} \\
&= S\mu \int_{Q_{\mathrm{S}}}^{Q_{\mathrm{D}}} \left(-\frac{Q_{\mathrm{m}}}{C} + U_{\mathrm{T}}\right) dQ_{\mathrm{m}} + S\mu \int_{Q_{\mathrm{D}}}^{0} \left(-\frac{Q_{\mathrm{m}}}{C} + U_{\mathrm{T}}\right) dQ_{\mathrm{m}} \\
&\qquad\qquad - S\mu \int_{Q_{\mathrm{D}}}^{0} \left(-\frac{Q_{\mathrm{m}}}{C} + U_{\mathrm{T}}\right) dQ_{\mathrm{m}} \\
&= \underbrace{S\mu \int_{Q_{\mathrm{S}}}^{0} \left(-\frac{Q_{\mathrm{m}}}{C} + U_{\mathrm{T}}\right) dQ_{\mathrm{m}}}_{I_{\mathrm{F}}} - \underbrace{S\mu \int_{Q_{\mathrm{D}}}^{0} \left(-\frac{Q_{\mathrm{m}}}{C} + U_{\mathrm{T}}\right) dQ_{\mathrm{m}}}_{I_{\mathrm{R}}}, \qquad (12)
\end{aligned}$$

where $S \equiv W/L$ is the *strength ratio* of the transistor, Q_{S} is the mobile charge per unit area at the source end of the channel and Q_{D} is the mobile charge per unit area at the drain end of the channel. Note that we have made use of Eq. 7 in going from the second step to the third step in the above derivation.

We have expressed the channel current as the difference between a forward current component, I_{F}, and a reverse current component, I_{R}, which have identical functional forms, except that I_{F} depends on Q_{S} and I_{R} depends on Q_{D}. The mobile charge density at the source end of the channel, in turn, will depend on the gate-to-bulk voltage and the source-to-bulk voltage and not on the drain-to-bulk voltage. In the same way, the mobile charge density at the drain end of the channel should depend on the gate-to-bulk potential and on the drain-to-bulk potential and not on the source-to-bulk potential. Moreover, we should expect that Q_{S} depends on V_{SB} in precisely the same way that Q_{D} depends on V_{DB}. Note that this MOS transistor model has the source/drain symmetric form expressed in Eq. 1. In such models, the primary channel current dependence on the drain voltage is contained wholly within I_{R}. If the forward current component is much larger than the reverse current component, then the channel current no longer depends significantly on the drain voltage, and the transistor is saturated. The saturation current is simply given by I_{F}. On the other hand, if the magnitudes of I_{F} and I_{R} are comparable, then the channel current depends strongly both on the source voltage and on the drain voltage. In this case, the transistor is in the ohmic region.

To make progress on the MOS transistor model beyond Eq. 12, we introduce an approximation, which was made initially by Maher and Mead [2, 3], that will allow us to evaluate the integrals in Eq. 12 in closed form. The channel capacitance per unit area, C, defined in Eq. 7, is a weak function of position along the length of the channel—in moderate and strong inversion, the depletion layer gets thicker closer to the drain end of the channel, making C_{dep} and, hence, C smaller nearer to the drain. We shall take C to be constant as a function of x, with C_{dep} replaced by an average value. With this approximation, we have that

$$\begin{aligned} I &= S\mu\left(-\frac{Q_{\mathrm{m}}^2}{2C}\bigg|_{Q_{\mathrm{S}}}^{0} + U_{\mathrm{T}}Q_{\mathrm{m}}\bigg|_{Q_{\mathrm{S}}}^{0}\right) - S\mu\left(-\frac{Q_{\mathrm{m}}^2}{2C}\bigg|_{Q_{\mathrm{D}}}^{0} + U_{\mathrm{T}}Q_{\mathrm{m}}\bigg|_{Q_{\mathrm{D}}}^{0}\right) \\ &= S\mu\left(\frac{Q_{\mathrm{S}}^2}{2C} - U_{\mathrm{T}}Q_{\mathrm{S}}\right) - S\mu\left(\frac{Q_{\mathrm{D}}^2}{2C} - U_{\mathrm{T}}Q_{\mathrm{D}}\right) \\ &= \underbrace{\frac{S\mu}{2C}\left(Q_{\mathrm{S}}^2 - 2CU_{\mathrm{T}}Q_{\mathrm{S}}\right)}_{I_{\mathrm{F}}} - \underbrace{\frac{S\mu}{2C}\left(Q_{\mathrm{D}}^2 - 2CU_{\mathrm{T}}Q_{\mathrm{D}}\right)}_{I_{\mathrm{R}}}. \end{aligned} \qquad (13)$$

Note that the quadratic terms in I_{F} and I_{R} in Eq. 13 originated with the drift component of the channel current, whereas the linear terms stemmed from the diffusion component. When $|Q_{\mathrm{S}}| \ll 2CU_{\mathrm{T}}$, the diffusion term in I_{F} is dominant over the drift term. Conversely, when $|Q_{\mathrm{S}}| \gg 2CU_{\mathrm{T}}$, then the drift term is much larger than the diffusion term. When $|Q_{\mathrm{S}}|$ is equal to $2CU_{\mathrm{T}}$, then these two current components are equal. This condition on the mobile charge density corresponds to the usual notion of *threshold*. Below threshold, the mobile charge density in the channel is small and the channel current is carried primarily by diffusion. In strong inversion, the mobile charge density is large, the channel current is carried mainly by drift. At threshold, the channel current flows both by drift and by diffusion. All of the same statements can also be made for $|Q_{\mathrm{D}}|$ and I_{R}. Thus, we shall take the magnitude of the mobile charge density at threshold to be given by $2CU_{\mathrm{T}}$.

Equation 13 represents a simple, closed-form expression for the channel current flowing in an nMOS transistor in terms of the mobile charge densities at the source and drain ends of the channel. This model equation is valid in all regions of normal MOS transistor operation, transistioning continuously from weak inversion to above threshold and from the ohmic region to the saturation region. Unfortunately, we would like to have the channel current explicitly in terms of the terminal voltages, V_{G}, V_{S}, and V_{D}. The dependence of the channel current on the terminal voltages come in through the dependence of Q_{S} on V_{G} and V_{S} and the dependence of Q_{D} on V_{G} and V_{D}. Unfortunately, no physically exact, closed-form expressions derived from first principles exist for these dependencies. In the next two sections, we shall explore two extreme limits of this model where simple approximate relationships between the mobile charge densities and the terminal voltages exist.

4.3 Weak-Inversion Operation

In the weak-inversion region of operation, the amount of mobile charge in the channel is negligible compared to the amount of charge exposed in the depletion layer beneath the gate. Thus, we should expect that the presence of the mobile charge will have a minuscule effect on the electrostatics in the channel region. If the electrostatics are only determined by the substrate potential, the gate potential, the gate charge, and the depletion charge, then, because the gate and the substrate are both isopotential, we would expect that the surface potential, ψ_{s}, should also be constant along the channel. Because the electric field is given by the gradient of the potential, a constant surface potential implies that there is no electric field in the direction of the channel. The absence of an electric field along the channel, in turn, implies that any current flow must be by diffusion rather than by drift, which is consistent with our conclusions at the end of Section 4.2 stemming from Eq. 13. Therefore, to obtain a weak-inversion model for MOS transistor operation, we should be able to neglect the quadratic terms in this equation, but we must still relate the mobile charge densities at the source and drain ends of the channel to the applied terminal voltages.

We know that, in weak inversion, there must be a relatively substantial energy barrier between the source and drain regions and the channel, otherwise there would be a substantial number of charges in the channel. The number of carriers in the source that have sufficient energy to surmount the energy barrier at the source end of the channel will follow the Boltzmann distribution, being exponential in the height of the energy barrier, which, in turn, is given by the difference between the source potential, V_{S}, and the surface potential, ψ_{s}. Likewise, the number of carriers in the drain that have sufficient energy to surmount the energy barrier at the drain end of the channel will be exponential in the difference between the V_{D} and ψ_{s}. Thus, we expect that

$$Q_{\mathrm{S}} \propto e^{(\psi_{\mathrm{s}}-V_{\mathrm{S}})/U_{\mathrm{T}}} \quad \text{and} \quad Q_{\mathrm{D}} \propto e^{(\psi_{\mathrm{s}}-V_{\mathrm{D}})/U_{\mathrm{T}}}.$$

Unfortunately, we need to have these charge densities in terms of the applied gate voltage and we will need to know what the constant of proportionally to use.

Toward this end, we shall again make use of Eq. 6 to determine how incremental changes in the gate voltage affect the surface potential. By neglecting the mobile charge term and differentiating with respect to ψ_{s}, we find that

$$\begin{aligned}\frac{\partial V_{\mathrm{G}}}{\partial \psi_{\mathrm{s}}} &= 1 - \frac{1}{C_{\mathrm{ox}}} \cdot \frac{\partial Q_{\mathrm{dep}}}{\partial \psi_{\mathrm{s}}} \\ &= 1 + \frac{C_{\mathrm{dep}}}{C_{\mathrm{ox}}} \\ &= \frac{C_{\mathrm{ox}} + C_{\mathrm{dep}}}{C_{\mathrm{ox}}},\end{aligned}$$

which implies that, in weak inversion, the incremental voltage gain from the gate to the surface is given by

$$\kappa \equiv \frac{\partial \psi_{\mathrm{s}}}{\partial V_{\mathrm{G}}} = \frac{C_{\mathrm{ox}}}{C_{\mathrm{ox}} + C_{\mathrm{dep}}}. \tag{14}$$

In this regime, the situation can be thought of intuitively as a capacitive voltage divider between the oxide capacitance above the channel and the effective capacitance of the depletion layer beneath the channel—this parameter is simply the capacitive divider ratio. Because the thickness of the depletion layer beneath the channel increases with increasing gate voltage, C_{dep} will get smaller for larger values of V_{G}, which implies that κ will increase with increasing V_{G}. However, κ is only a slowly-varying function of V_{G}, and even for moderate changes in V_{G}, it is reasonable to assume that the value of κ is constant with a value between 0.5 and 0.9.

For relatively small changes about some operating point, we can expand ψ_{s} in a Taylor series and truncate after the linear term. In order to also get the constant of proportionality in the mobile charge densities, we shall choose to expand the surface potential around the point $V_{\mathrm{G}} = V_{\mathrm{T0}}$, the *zero-bias threshold voltage*. When the gate voltage is equal to V_{T0} with the source and drain grounded, Q_{S} and Q_{D} should both be equal to $-2CU_{\mathrm{T}}$. Thus, we have that the mobile charge densities at the source and drain end of the channel are given approximately by

$$Q_{\mathrm{S}} \approx -2CU_{\mathrm{T}} e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/U_{\mathrm{T}}} \quad \text{and} \quad Q_{\mathrm{D}} \approx -2CU_{\mathrm{T}} e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{D}})/U_{\mathrm{T}}}. \tag{15}$$

Retaining only the terms in Eq. 13 that stem from diffusion and substituting the mobile charge densities given in Eq. 15, we find that the channel current in an nMOS transistor operating in weak inversion is given by

$$\begin{aligned}
I &= S\mu U_{\mathrm{T}}\left(-Q_{\mathrm{S}}\right) - S\mu U_{\mathrm{T}}\left(-Q_{\mathrm{D}}\right) \\
&= 2S\mu C U_{\mathrm{T}}^2 e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/U_{\mathrm{T}}} - 2S\mu C U_{\mathrm{T}}^2 e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{D}})/U_{\mathrm{T}}} \\
&= 2S\mu C U_{\mathrm{T}}^2 e^{\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})/U_{\mathrm{T}}} \left(e^{-V_{\mathrm{S}}/U_{\mathrm{T}}} - e^{-V_{\mathrm{D}}/U_{\mathrm{T}}}\right) \\
&= \frac{2S\mu C_{\mathrm{ox}} U_{\mathrm{T}}^2}{\kappa} e^{\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})/U_{\mathrm{T}}} \left(e^{-V_{\mathrm{S}}/U_{\mathrm{T}}} - e^{-V_{\mathrm{D}}/U_{\mathrm{T}}}\right) \\
&= I_{\mathrm{s}} e^{\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})/U_{\mathrm{T}}} \left(e^{-V_{\mathrm{S}}/U_{\mathrm{T}}} - e^{-V_{\mathrm{D}}/U_{\mathrm{T}}}\right),
\end{aligned} \tag{16}$$

where we have used Eq. 14 to express C as C_{ox}/κ and we have introduced

$$I_{\mathrm{s}} \equiv \frac{2S\mu C_{\mathrm{ox}} U_{\mathrm{T}}^2}{\kappa},$$

which corresponds to approximately twice the threshold current of the transistor.

Equation 16 represents a complete model for the operation of an nMOS transistor in weak inversion, covering both the ohmic region and the saturation region. To see that it does, we can rearrange Eq. 16 to obtain

$$\begin{aligned}
I &= I_{\mathrm{s}} e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/U_{\mathrm{T}}} \left(1 - e^{-V_{\mathrm{DS}}/U_{\mathrm{T}}}\right) \\
&= I_{\mathrm{sat}} \left(1 - e^{-V_{\mathrm{DS}}/U_{\mathrm{T}}}\right) \\
&\approx I_{\mathrm{sat}}
\end{aligned}$$

when V_{DS} is larger than about $4U_{\mathrm{T}}$ or $5U_{\mathrm{T}}$. Thus, the channel current saturates, becoming independent of V_{DS} for $V_{\mathrm{DS}} \geq 5U_{\mathrm{T}}$. For small V_{DS}, to see that the model predicts ohmic behavior, we can expand the $e^{-V_{\mathrm{DS}}/U_{\mathrm{T}}}$ term in this equation in a Taylor series around $V_{\mathrm{DS}} = 0$,

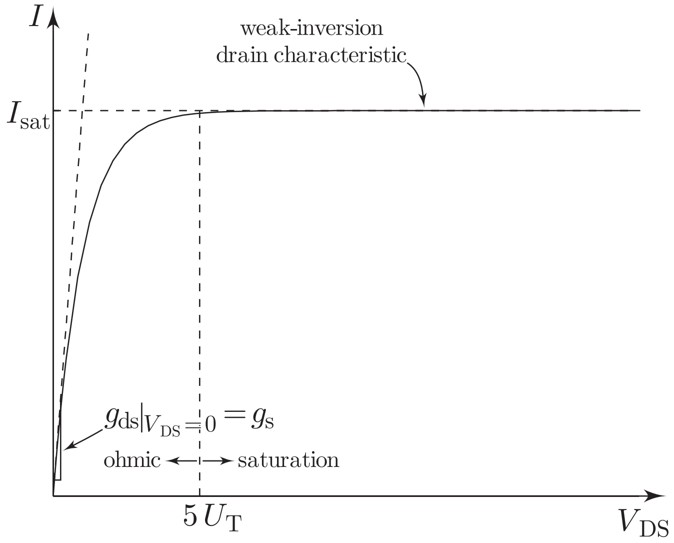


Figure 2: Typical weak-inversion drain characteristic.

retaining only the linear term. Doing so, we find that

$$\begin{aligned} I &= I_{\text{sat}}\left(1 - e^{-V_{\text{DS}}/U_{\text{T}}}\right) \\ &= I_{\text{sat}}\left(1 - \left(1 - \frac{V_{\text{DS}}}{U_{\text{T}}} + \frac{1}{2}\left(\frac{V_{\text{DS}}}{U_{\text{T}}}\right)^2 - \cdots\right)\right) \\ &\approx I_{\text{sat}}\left(1 - 1 + \frac{V_{\text{DS}}}{U_{\text{T}}}\right) \\ &= \frac{I_{\text{sat}}}{U_{\text{T}}} \cdot V_{\text{DS}} \\ &= g_{\text{ds}} V_{\text{DS}}, \end{aligned}$$

where $g_{\text{ds}} \equiv I_{\text{sat}}/U_{\text{T}}$ represents the incremental conductance of the channel deep in the ohmic region. These behaviors are both evident in a typical weak-inversion drain characteristic, as shown in Fig. 2.

The weak-inversion model given in Eq. 16 is formally identical to the ones presented by Mead [4], Vittoz [5], and Bult [6].

4.4 Strong-Inversion Operation

In strong inversion, the mobile charge density in the channel exceeds the charge density in the depletion layer, and the channel charge has a large effect on the surface potential along the channel. As discussed at the end of Section 4.2, for the levels of mobile charge density that occur in strong inversion, the current flow is predominantly by drift. Consequently, in developing a strong-inversion model for the MOS transistor, we shall only retain the quadratic terms in Eq. 13. In this regime, the energy barriers at the source and drain ends of the channel have been reduced to such an extent that nearly every additional charge that we place on the gate is balanced by additional mobile charges in the channel rather than by uncovering more charge in the depletion layer beneath the channel. Moreover, the inversion layer basically serves as the bottom plate of a parallel plate capacitor between the gate and the channel whose capacitance per unit area is just C_{ox}.

To obtain expressions for the mobile charge densities at the source and drain end of the channel, we shall assume that all of the gate charges that go into raising the gate voltage up to the threshold voltage are balanced by fixed charges in the depletion layer beneath the channel and that all of the gate charges that go into raising the gate voltage above the threshold voltage are balanced by additional mobile charges in the channel. Thus, we have that the mobile charge densities in strong inversion are given by

$$Q_{\mathrm{S}} = -C_{\mathrm{ox}}\left(V_{\mathrm{G}} - V_{\mathrm{T}}(V_{\mathrm{S}})\right) \quad \text{and} \quad Q_{\mathrm{D}} = -C_{\mathrm{ox}}\left(V_{\mathrm{G}} - V_{\mathrm{T}}(V_{\mathrm{D}})\right), \tag{17}$$

where $V_{\mathrm{T}}(V_{\mathrm{S}})$ and $V_{\mathrm{T}}(V_{\mathrm{D}})$ are the (bulk-referenced) threshold voltages at the source and drain ends of the channel, respectively. The threshold voltage at the source end of the channel represents the gate voltage that we need to apply given the source voltage in order for the mobile charge density at the source end of the channel to just equal $-2CU_{\mathrm{T}}$. We can obtain a simple approximate expression for $V_{\mathrm{T}}(V_{\mathrm{S}})$ using the weak-inversion expression for Q_{S} given in Eq. 15 by determining the value of V_{G} that makes Q_{S} equal to $-2CU_{\mathrm{T}}$. Doing so, we find that

$$V_{\mathrm{T}}(V_{\mathrm{S}}) = V_{\mathrm{T0}} + \frac{V_{\mathrm{S}}}{\kappa}. \tag{18}$$

Note that we are here taking into account the threshold-voltage increase normally associated with the *body effect* using a simple linear approximation, similar to the one introduced by Wallinga and Bult [7]. Thus, the strong-inversion model that we are developing accounts directly for the body effect (to first order) via the κ parameter without any auxiliary equations and without adding too much complexity to the model. Similarly, the threshold voltage at the drain end of the channel is the gate voltage that we need to apply given the drain voltage so that the Q_{D} is just equal to $-2CU_{\mathrm{T}}$. Using the same approach as we just took for $V_{\mathrm{T}}(V_{\mathrm{S}})$, we find that

$$V_{\mathrm{T}}(V_{\mathrm{D}}) = V_{\mathrm{T0}} + \frac{V_{\mathrm{D}}}{\kappa}. \tag{19}$$

Retaining only the terms in Eq. 13 that stem from drift and substituting the mobile charge densities and threshold voltages given in Eqs. 17, 18, and 19, we find that the channel current in an *n*MOS transistor operating in strong inversion is given by

$$\begin{aligned} I &= \frac{S\mu}{2C}Q_{\mathrm{S}}^2 - \frac{S\mu}{2C}Q_{\mathrm{D}}^2 \\ &= \frac{S\mu}{2C}\left(\left(-C_{\mathrm{ox}}\left(V_{\mathrm{G}} - V_{\mathrm{T0}} - \frac{V_{\mathrm{S}}}{\kappa}\right)\right)^2 - \left(-C_{\mathrm{ox}}\left(V_{\mathrm{G}} - V_{\mathrm{T0}} - \frac{V_{\mathrm{D}}}{\kappa}\right)\right)^2\right) \\ &= \frac{S\mu}{2C}\cdot\frac{C_{\mathrm{ox}}^2}{\kappa^2}\left(\left(\kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{S}}\right)^2 - \left(\kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{D}}\right)^2\right) \\ &= \frac{S\mu C_{\mathrm{ox}}}{2\kappa}\left(\left(\kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{S}}\right)^2 - \left(\kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{D}}\right)^2\right). \end{aligned} \tag{20}$$

Unfortunately, Eq. 20 only captures the behavior of the MOS transistor in the ohmic region in strong inversion. To see that it does, we rearrange Eq. 20 slightly to obtain

$$I = \frac{S\mu C_{\mathrm{ox}}}{2\kappa}\left(\left(\kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{S}}\right)^2 - \left(\kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{S}} + V_{\mathrm{S}} - V_{\mathrm{D}}\right)^2\right)$$

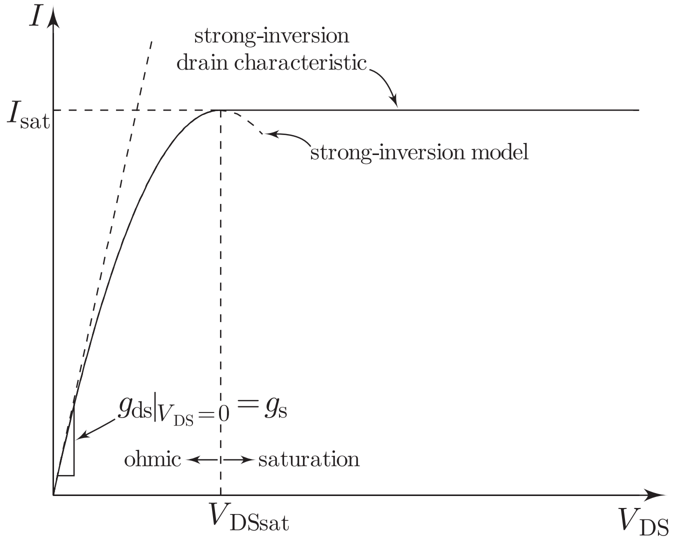


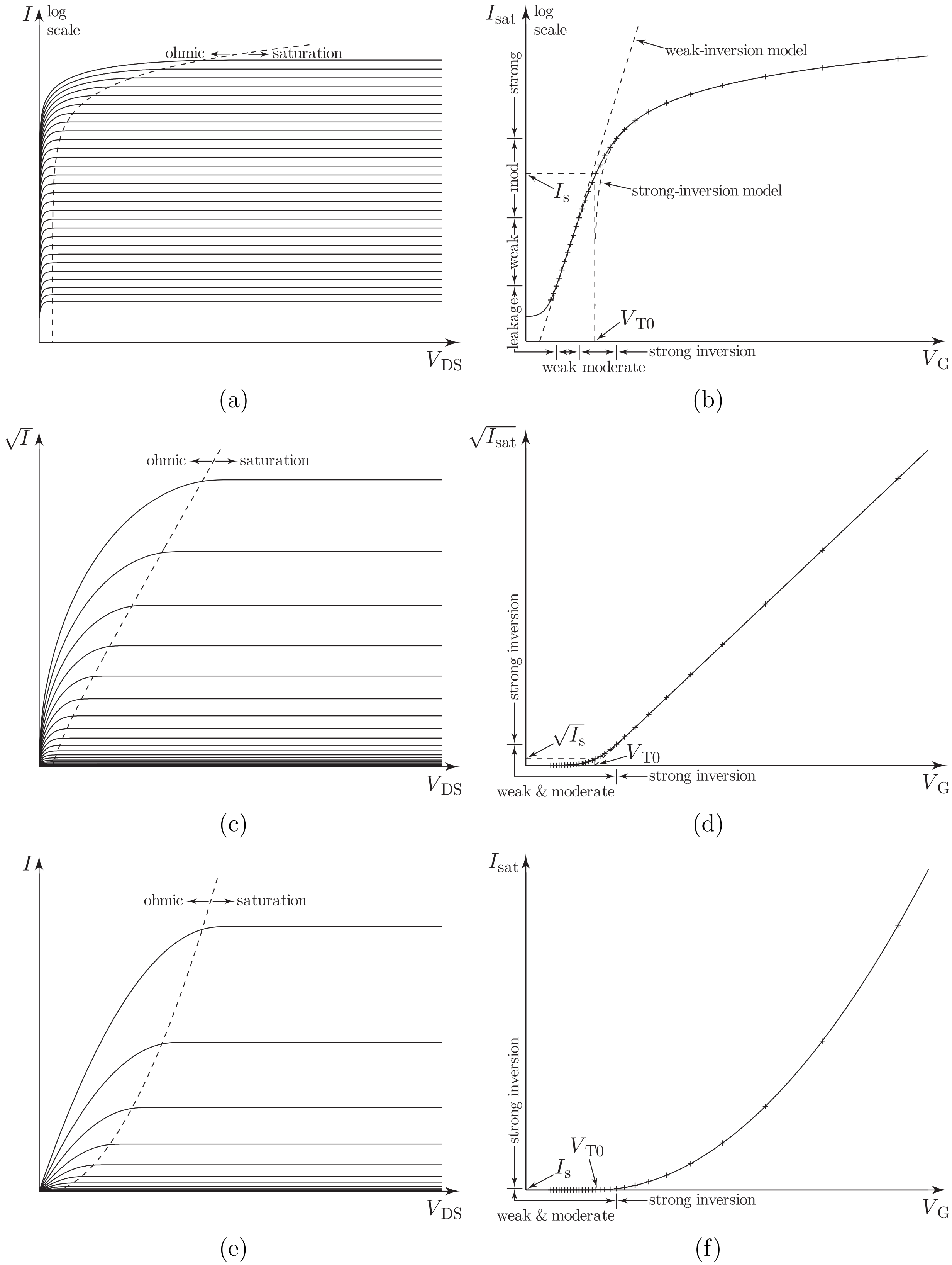
Figure 3: Typical strong-inversion drain characteristic.

$$\begin{aligned}
&= \frac{S\mu C_{\rm ox}}{2\kappa}\left(\left(\kappa\left(V_{\rm G}-V_{\rm T0}\right)-V_{\rm S}\right)^2-\left(\kappa\left(V_{\rm G}-V_{\rm T0}\right)-V_{\rm S}-V_{\rm DS}\right)^2\right)\\
&= \frac{S\mu C_{\rm ox}}{2\kappa}\left(\kappa\left(V_{\rm G}-V_{\rm T0}\right)-V_{\rm S}\right)^2\left(1-\left(1-\frac{V_{\rm DS}}{\kappa\left(V_{\rm G}-V_{\rm T0}\right)-V_{\rm S}}\right)^2\right)\\
&= \frac{S\mu C_{\rm ox}}{2\kappa}V_{\rm DSsat}^2\left(1-\left(1-\frac{V_{\rm DS}}{V_{\rm DSsat}}\right)^2\right)\\
&= I_{\rm sat}\left(1-\left(1-\frac{V_{\rm DS}}{V_{\rm DSsat}}\right)^2\right), \qquad (21)
\end{aligned}$$

where $V_{\rm DSsat} \equiv \kappa\left(V_{\rm G}-V_{\rm T0}\right)-V_{\rm S}$ and

$$I_{\rm sat} \equiv \frac{S\mu C_{\rm ox}}{2\kappa}V_{\rm DSsat}^2.$$

Figure 3 shows a typical strong-inversion drain characteristic along with a plot of Eq. 21. The two curves agree for $V_{\rm DS} \leq V_{\rm DSsat}$, but the model equation turns around while the drain characteristic saturates for $V_{\rm DS} > V_{\rm DSsat}$. This deviation occurs because at $V_{\rm DS} = V_{\rm DSsat}$, the value of $Q_{\rm D}$ given in Eq. 17 is equal to zero and for $V_{\rm DS} > V_{\rm DSsat}$, it would become *positive*, which does not happen physically in this region of operation because the mobile charges in the channel are electrons. Rather, what actually happens is that the drain end of the channel goes into weak inversion and $Q_{\rm D}$ transitions smoothly over to the exponential form given in Eq. 15. Understanding that the second term in Eq. 20 comes from $Q_{\rm D}$ makes it clear how the transistor transitions from the ohmic region to the saturation region in strong inversion. In saturation, the first term in Eq. 20 is much larger than the second one, so we can neglect the second one and the current is described by the first one by itself. This point is almost completely obscured by conventional source-referenced models of the strong-inversion MOS transistor. The model given in Eq. 20 is formally identical to the one given by Wallinga and Bult [7] and by Vittoz [5].



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4: MOS-transistor current–voltage characteristics plotted on various scales.

4.5 A Complete Set of Characteristics

Figure 4 shows a complete set of MOS-transistor current–voltage characteristics plotted on various scales to highlight the different regions of operation. The three plots on the left in Fig. 4 show a family of drain characteristics for saturation currents equally spaced on a log scale ranging from weak inversion through moderate inversion to strong inversion. These plots also show the onset of saturation (i.e., V_{DSsat}) as a dashed line. Note that, in weak inversion, the saturation voltage is independent of the current level, but begins to increase with increasing current level as the transistor enters strong inversion. The three plots on the right show channel current in saturation as a function of gate voltage. The points indicated with plus point markers show the set of gate voltages and saturation currents that correspond to the drain characteristics in the plots on the left.

4.6 The EKV Model

The Enz-Krummenacher-Vittoz (EKV) model [5,8] of the MOS transistor provides a simple approximate closed-form expression for the channel current of a MOS transistor in terms of the terminal voltages, each of which is referenced to the transistor's bulk voltage. It is valid in all regions of normal MOS transistor operation (i.e., when the drain-bulk and the source-bulk junctions are reversed biased), transitioning between them continuously. In its simplest form, the EKV model expresses the channel current in an nMOS transistor as

$$I = \underbrace{I_{\mathrm{s}} \log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right)}_{I_{\mathrm{F}}} - \underbrace{I_{\mathrm{s}} \log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{D}})/2U_{\mathrm{T}}}\right)}_{I_{\mathrm{R}}}, \qquad (22)$$

where all of the terms have their previously defined meanings.

The function $\log^2\left(1 + e^{x/2}\right)$ interpolates smoothly between an exponential (i.e., weak-inversion behavior) when $x < 0$ and a quadratic (i.e., strong-inversion behavior) when $x > 0$. To see that it does, first we suppose that $x < 0$. Then, it follows that $e^{x/2} \lll 1$ and we have that $\log\left(1 + e^{x/2}\right) \approx e^{x/2}$, because $\log(1 + y) \approx y$ for small $|y| \lll 1$. Finally, because $\left(e^{x/2}\right)^2 = e^x$, we have that $\log^2\left(1 + e^{x/2}\right) \approx e^x$ for $x < 0$. Conversely, suppose that $x > 0$. Then, it follows that $e^{x/2} \ggg 1$ and $1 + e^{x/2} \approx e^{x/2}$. Finally, because $\log e^{x/2} = x/2$, we have that $\log^2\left(1 + e^{x/2}\right) \approx (x/2)^2$ for $x > 0$.

Consequently, if both $V_{\mathrm{G}} < V_{\mathrm{T0}} + V_{\mathrm{S}}/\kappa$ and $V_{\mathrm{G}} < V_{\mathrm{T0}} + V_{\mathrm{D}}/\kappa$, then Eq. 22 becomes

$$\begin{aligned} I &= I_{\mathrm{s}} \left(\log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) - \log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{D}})/2U_{\mathrm{T}}}\right)\right) \\ &\approx I_{\mathrm{s}} \left(e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/U_{\mathrm{T}}} - e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{D}})/U_{\mathrm{T}}}\right) \\ &= I_{\mathrm{s}} e^{\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}})/U_{\mathrm{T}}} \left(e^{-V_{\mathrm{S}}/U_{\mathrm{T}}} - e^{-V_{\mathrm{D}}/U_{\mathrm{T}}}\right), \end{aligned}$$

and we recover the weak-inversion model given in Eq. 16. Conversely, if both $V_{\mathrm{G}} > V_{\mathrm{T0}} + V_{\mathrm{S}}/\kappa$ and $V_{\mathrm{G}} > V_{\mathrm{T0}} + V_{\mathrm{D}}/\kappa$, then Eq. 22 becomes

$$I = I_{\mathrm{s}} \left(\log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) - \log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{D}})/2U_{\mathrm{T}}}\right)\right)$$

$$
\begin{aligned}
&\approx I_{\mathrm{s}}\left(\left(\frac{\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{S}}}{2U_{\mathrm{T}}}\right)^{2}-\left(\frac{\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{D}}}{2U_{\mathrm{T}}}\right)^{2}\right)\\
&= \frac{I_{\mathrm{s}}}{4U_{\mathrm{T}}^{2}}\left(\left(\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{S}}\right)^{2}-\left(\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{D}}\right)^{2}\right)\\
&= \frac{S\mu C_{\mathrm{ox}}}{2\kappa}\left(\left(\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{S}}\right)^{2}-\left(\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{D}}\right)^{2}\right),
\end{aligned}
$$

and we recover the strong-inversion model given in Eq. 20. Note that, if $V_{\mathrm{G}} > V_{\mathrm{T0}} + V_{\mathrm{S}}/\kappa$ but $V_{\mathrm{G}} < V_{\mathrm{T0}} + V_{\mathrm{D}}/\kappa$, then Eq. 22 becomes

$$
\begin{aligned}
I &= I_{\mathrm{s}}\left(\log^{2}\left(1+e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right)-\log^{2}\left(1+e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{D}})/2U_{\mathrm{T}}}\right)\right)\\
&\approx I_{\mathrm{s}}\left(\left(\frac{\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{S}}}{2U_{\mathrm{T}}}\right)^{2}-e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{D}})/U_{\mathrm{T}}}\right)\\
&\approx \frac{I_{\mathrm{s}}}{4U_{\mathrm{T}}^{2}}\left(\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{S}}\right)^{2}\\
&= \frac{S\mu C_{\mathrm{ox}}}{2\kappa}\left(\kappa\left(V_{\mathrm{G}}-V_{\mathrm{T0}}\right)-V_{\mathrm{S}}\right)^{2},
\end{aligned}
$$

which is just the strong-inversion model in the saturation region. The EKV model switches between these forms smoothly with no discontinuities.

Figure 4b shows a semilog plot of the saturation currents predicted by the weak-inversion model, the strong-inversion model, and a typical MOS-transistor current–voltage characteristic. Note that both the weak-inversion model and the strong-inversion model deviate substantially from actual MOS-transistor behavior for $V_{\mathrm{G}} \approx V_{\mathrm{T0}}$, which indicates that, if we were to assume either of these two models in the moderate-inversion region, our conclusions based on such an assumption would be highly suspect. The simple EKV model represents an excellent tool for reasoning about CMOS circuits in all regions of operation, *including* moderate inversion, which is becoming increasingly important for CMOS circuit design.

4.7 The Onset of Saturation

In Section 4.2, we noted that, if the forward component of the channel current, which according to the EKV model is given by

$$
I_{\mathrm{F}} = I_{\mathrm{s}}\log^{2}\left(1+e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right),
$$

is much larger than the reverse one, which is given by

$$
\begin{aligned}
I_{\mathrm{R}} &= I_{\mathrm{s}}\log^{2}\left(1+e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{D}})/2U_{\mathrm{T}}}\right)\\
&= I_{\mathrm{s}}\log^{2}\left(1+e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}}-V_{\mathrm{DS}})/2U_{\mathrm{T}}}\right)\\
&= I_{\mathrm{s}}\log^{2}\left(1+e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/2U_{\mathrm{T}}}e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}\right),
\end{aligned}
$$

then the channel current no longer depends significantly on the drain voltage, which corresponds to the saturation region of operation. In this notion, we find the basis for a generic definition of the onset of saturation for all levels of inversion in terms of an arbitrary parameter, A, that is useful from a circuit-design standpoint. We will say that an MOS transistor is saturated if and only if $I_{\mathrm{F}}/I_{\mathrm{R}} \geq A$, where $A \gg 1$. The saturation voltage, V_{DSsat}, would then be given by the value of V_{DS} that makes $I_{\mathrm{F}}/I_{\mathrm{R}}$ equal to A.

To find an explicit expression for V_{DSsat}, we set the ratio of I_{F} to I_{R} equal to A and solve for V_{DSsat}. Doing so, we write that

$$A = \frac{I_{\mathrm{F}}}{I_{\mathrm{R}}} = \frac{\log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right)}{\log^2\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}} - V_{\mathrm{DSsat}})/2U_{\mathrm{T}}}\right)},$$

which we can invert as follows. By rearranging this equation and taking the square root of both sides, we find that

$$\begin{aligned}\log\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}} e^{-V_{\mathrm{DSsat}}/2U_{\mathrm{T}}}\right) &= \frac{1}{\sqrt{A}} \log\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\ &= \log\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right)^{1/\sqrt{A}}.\end{aligned}$$

By exponentiating both sides of this equation, we obtain

$$1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}} e^{-V_{\mathrm{DSsat}}/2U_{\mathrm{T}}} = \left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right)^{1/\sqrt{A}},$$

which we can rearrange to find that

$$e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}} e^{-V_{\mathrm{DSsat}}/2U_{\mathrm{T}}} = \left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right)^{1/\sqrt{A}} - 1.$$

Rearranging this equation to isolate V_{DSsat}, we get

$$e^{V_{\mathrm{DSsat}}/2U_{\mathrm{T}}} = \frac{e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}}{(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}})^{1/\sqrt{A}} - 1}.$$

Finally, by taking the natural logarithm of both sides of this equation, we find that V_{DSsat} is given by

$$\begin{aligned} V_{\mathrm{DSsat}} &= \kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{S}} - 2U_{\mathrm{T}} \log\left(\left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right)^{1/\sqrt{A}} - 1\right) \qquad (23) \\ &\approx \begin{cases} U_{\mathrm{T}} \log A, & V_{\mathrm{G}} < V_{\mathrm{T0}} + \dfrac{V_{\mathrm{S}}}{\kappa} \\ \left(1 - \dfrac{1}{\sqrt{A}}\right)\left(\kappa\left(V_{\mathrm{G}} - V_{\mathrm{T0}}\right) - V_{\mathrm{S}}\right), & V_{\mathrm{G}} > V_{\mathrm{T0}} + \dfrac{V_{\mathrm{S}}}{\kappa}. \end{cases}\end{aligned}$$

In both the weak-inversion and strong-inversion cases, the dependence of V_{DSsat} on A is only a weak one, being logarithmic in weak inversion and square-root in strong inversion. Thus, the our choice of A is not critical. By setting A equal to 100, the saturation voltage in weak inversion predicted by this expression is approximately $4.6U_{\mathrm{T}}$, which correlates well with the value of $4U_{\mathrm{T}}$ or $5U_{\mathrm{T}}$ that we identified for V_{DSsat} in Section 4.3. Moreover, the strong-inversion

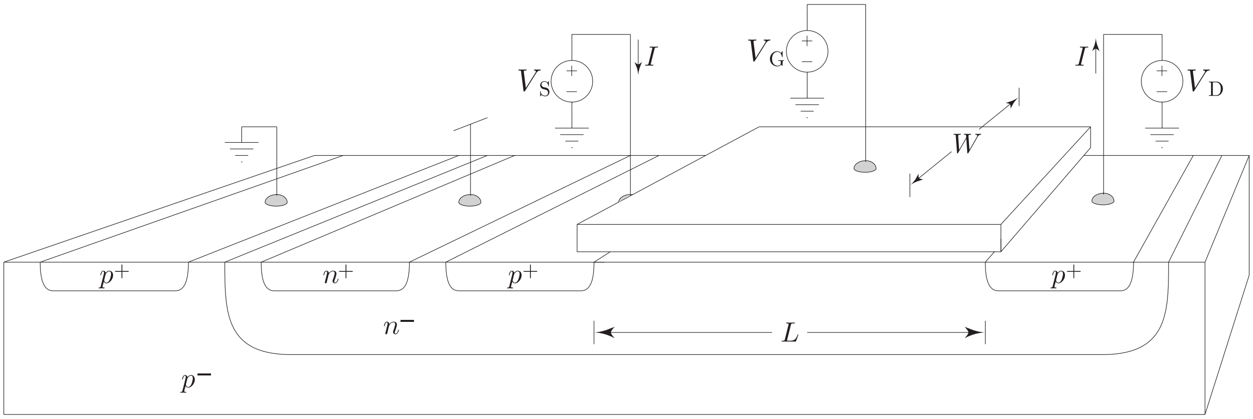


Figure 5: Simplified structure of a *p*MOS transistor fabricated in an *n*-well technology.

expression for V_{DSsat} approaches the one that we defined in Section 4.4 asymptotically as A becomes large.

We can also obtain an expression for V_{DSsat} in terms of the saturation current, I_{sat}, by noting that

$$I_{\mathrm{sat}} \approx I_{\mathrm{F}} = I_{\mathrm{s}} \log^2\left(1 + e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right),$$

which implies both that

$$\log\left(1 + e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) = \sqrt{\frac{I_{\mathrm{sat}}}{I_{\mathrm{s}}}} \tag{24}$$

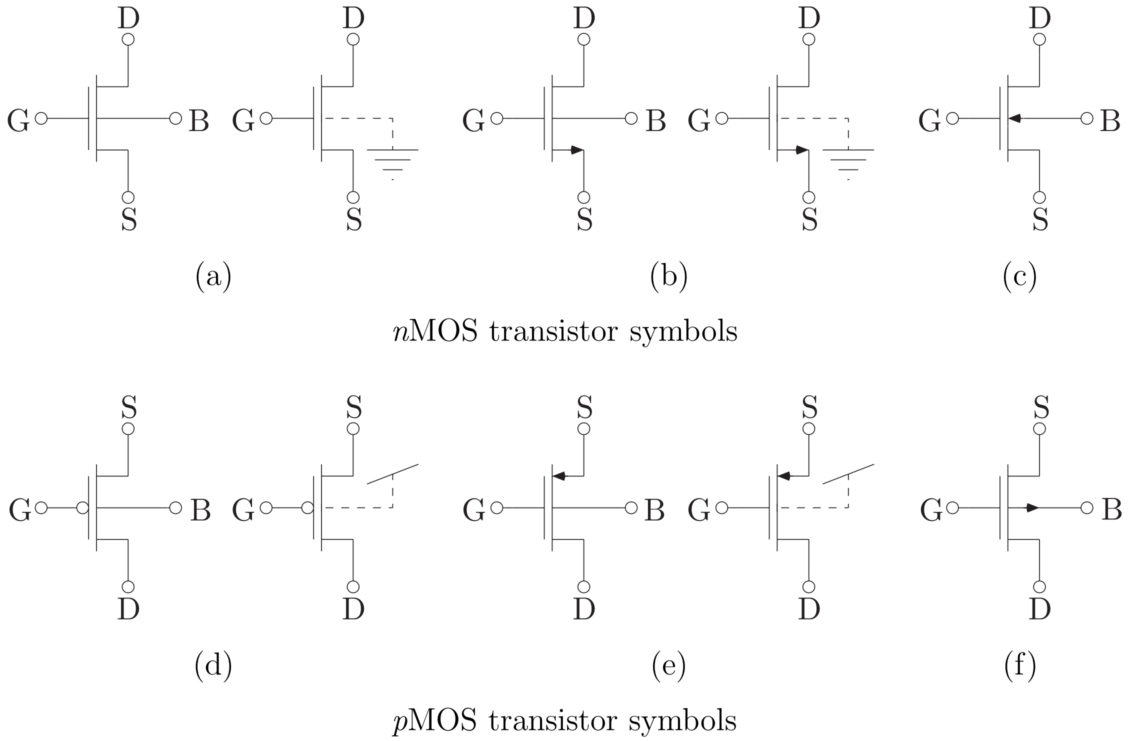
and that

$$e^{(\kappa(V_{\mathrm{G}}-V_{\mathrm{T0}})-V_{\mathrm{S}})/2U_{\mathrm{T}}} = e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}} - 1. \tag{25}$$

By substituting Eq. 24 and Eq. 25 into Eq. 23, we can express V_{DSsat} in terms of I_{sat}, as

$$\begin{aligned} V_{\mathrm{DSsat}} &= 2U_{\mathrm{T}} \log\left(e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}} - 1\right) - 2U_{\mathrm{T}} \log\left(e^{\sqrt{I_{\mathrm{sat}}/AI_{\mathrm{s}}}} - 1\right) \qquad (26) \\ &\approx \begin{cases} U_{\mathrm{T}} \log A, & I_{\mathrm{sat}} \ll I_{\mathrm{s}} \\ 2U_{\mathrm{T}}\left(1 - \dfrac{1}{\sqrt{A}}\right)\sqrt{\dfrac{I_{\mathrm{sat}}}{I_{\mathrm{s}}}}, & I_{\mathrm{sat}} \gg I_{\mathrm{s}}. \end{cases} \end{aligned}$$

We note that this formulation of the saturation voltage is similar to one introduced by Montoro and Cunha [9]. They have chosen to define the onset of saturation in terms of ratio of the mobile charge densities at the source and drain ends of the channel, rather than in terms of the ratio of the magnitudes of the forward and reverse current components. It seems that the formulation in terms of $I_{\mathrm{F}}/I_{\mathrm{R}}$ is more useful from a circuit-design viewpoint, because we can set up these current components directly using transistor strength ratios and principles such as the MOS current-division principle [10] in a way that is theoretically independent of the current level in the transistors.



*n*MOS transistor symbols

*p*MOS transistor symbols

Figure 6: Commonly used MOS transistor symbols.

4.8 The *p*MOS Transistor

Figure 5 depicts the structure of a *p*MOS transistor fabricated in an *n*-well CMOS technology. The source and drain regions of the *p*MOS transistor are heavily doped *p*-type regions, separated from each other by a shallow lightly doped *n*-type well region that has been formed at the surface of the *p*-type substrate. In order to keep the drain-bulk and source-bulk *p*-*n* junctions reverse biased, we must ensure that the drain and source voltages are always less than or equal to the well voltage, which is often connected to the positive power supply, V_{DD}. The behavior of the *p*MOS transistor is complementary to that of the *n*MOS transistor. The channel current is carried by holes rather than by electrons and flows from source to drain rather than from drain to source. The magnitude of the channel current increases as the applied gate voltage decreases or as the applied source voltage increases.

If we take the threshold voltage of the *p*MOS transistor to be a positive number, we can use all of the model equations that we developed for the *n*MOS transistor for the *p*MOS transistor by simply replacing V_{G} with $V_{\mathrm{W}} - V_{\mathrm{G}}$, V_{S} with $V_{\mathrm{W}} - V_{\mathrm{S}}$, and V_{D} with $V_{\mathrm{W}} - V_{\mathrm{D}}$, where V_{W} is the well voltage. For example, if the well were connected to V_{DD}, as shown in Fig. 5, we the channel current, I, would be given by

$$\begin{aligned} I \;&=\; I_{\mathrm{s}}\Big(\log^2\Big(1 + e^{(\kappa((V_{\mathrm{DD}}-V_{\mathrm{G}})-V_{\mathrm{T0}})-(V_{\mathrm{DD}}-V_{\mathrm{S}}))/2U_{\mathrm{T}}}\Big) \\ &\qquad\qquad - \log^2\Big(1 + e^{(\kappa((V_{\mathrm{DD}}-V_{\mathrm{G}})-V_{\mathrm{T0}})-(V_{\mathrm{DD}}-V_{\mathrm{D}}))/2U_{\mathrm{T}}}\Big)\Big). \end{aligned}$$

4.9 MOS Transistor Symbols

Figure 6 show various circuit symbols that are in common use to represent MOS transistors in circuit schematics. The *n*MOS symbols shown in Fig. 6a, Fig. 6b, and Fig. 6c, correspond to the *p*MOS symbols shown in Fig. 6d, Fig. 6e, and Fig. 6f, respectively. The symbols of Fig. 6a and Fig. 6d are used primarily by digital circuit designers while those of Fig. 6b and Fig. 6e are often used by analog circuit designers. We shall use the symbols of Fig. 6a and Fig. 6d both for analog circuits and for digital ones. The symbols shown in Fig. 6c and Fig. 6f are used primarily by device physicists who do not regularly have to construct or to understand complex circuit schematics. While MOS transistors are four-terminal devices, the bulk terminals of most transistors on integrated circuits are tied to one of the two power supply rails. For *n*MOS transistors, the bulk is usually connected to ground, whereas for *p*MOS transistors, the bulk is usually tied to V_{DD}. In order to simplify circuit schematics, we usually suppress drawing the bulk connections if they are connected in this way, as indicated by the three terminal variants shown in Fig. 6a, Fig. 6b, Fig. 6d, and Fig. 6e. We cannot use this convention with the symbols of Fig. 6c and Fig. 6f, because the direction of the arrow on the bulk terminal provides the only visual distinction between the *n*MOS symbol and the *p*MOS one.

The symbols shown in Fig. 6a, Fig. 6c, Fig. 6d, and Fig. 6f all emphasize the functional symmetry of the source and drain terminals, whereas the symbols shown in Fig. 6b and Fig. 6e indicate clearly the location of the source terminal of each MOS transistor. While this direct visual indication of the source terminal's location may initially seem to ease the difficulty inherent in parsing a complex circuit schematic, we maintain that these symbols have several undesirable features that more than offset this often cited "benefit."

First, in circuit symbols for semiconductor devices, arrows almost invariably indicate the presence of a *p*-*n* junction, with the arrow pointing from the *p*-type side of the junction to the *n*-type side. While it is true that the source-bulk junction of each MOS transistor is a *p*-*n* junction with the orientation indicated by the source arrow in these symbols, the source-bulk junction is always reverse-biased in normal MOS-transistor operation. Having such an arrow on the source terminal gives the indication that there is a *p*-*n* junction that becomes forward biased when the devices conducts a current, as would be the case for a diode or a bipolar transistor. However, in an MOS transistor, the surface is *inverted* to a greater or lesser extent when the device conducts a current and the charges that conduct the current are *majority* carriers in the inverted channel region rather than minority ones, as they would be in the base region of a bipolar transistor. Second, the symbols of Fig. 6b and Fig. 6e are not really distinct enough for us to be able to tell at a single glance which transistors in a schematic are *n*MOS and which are *p*MOS. The presence or absence of a bubble on the gate provides us with a much greater visual distinction than does the direction of the arrow on the source terminal.

Finally, most MOS transistors are, in fact, source/drain symmetric devices. In many circuits, which of the source/drain terminals acts as the source and which acts as the drain changes with time as the circuit operates. For an *n*MOS transistor, whichever of the source/drain terminals is at a higher potential serves as the drain and whichever is at a lower potential serves as the source. For a *p*MOS transistor, whichever of the source/drain terminals is at a higher potential serves as the source and whichever is at a lower potential serves as the drain. If the potentials applied to the source/drain terminals of either type of device are interchanged, the same channel current flows, only in the opposite direction. In circuits where the potential across an MOS transistor reverses polarity during normal operation, choosing one of these terminals a priori to be the "source" and labelling it as such gives an incorrect indication of the actual location of source terminal at some points in time. In some cases, for such MOS transistors, some people opt to use the symbols of Fig. 6c and Fig. 6f and those of Fig. 6b and Fig. 6e for those transistors whose source terminal remains invariant. While this practice is at least not misleading, it is rather cumbersome compared to using the symbols of Fig. 6a and Fig. 6d uniformly. Moreover, we normally construct circuit schematics so that nodes with higher potentials appear higher on the page and so that current flows down the page. In this case, for a vertically oriented *n*MOS transistor symbol, the source will be the terminal nearest the bottom of the page and, for a vertically oriented *p*MOS transistor, the source will be the terminal nearest the top of the page. If this is the case, then indicating the location of the source terminal with an arrow is somewhat redundant. MOS transistors whose source and drain terminals may interchange during circuit operation are often oriented *horizontally* in circuit schematics.

While the symbols of Fig. 6c and Fig. 6f are source/drain symmetric, they are awkward compared with those of Fig. 6a and Fig. 6d. In these symbols, the arrow indicates the presence and polarity of a so-called field-induced *p*-*n* junction that exists between the bulk and the channel, when the surface is (strongly) inverted. For an *n*MOS transistor, the bulk is *p*-type and the inverted channel region is *n*-type, so the arrow points towards the channel. For a *p*MOS transistor, the bulk is *n*-type and the inverted channel region is *p*-type, so the arrow points towards the bulk terminal. However, during normal MOS-transistor operation, no current flows from the channel to the bulk, because this "junction" is always reverse biased. While this "junction" is arguably present when the device is "on," it disappears when the device is "off." Moreover, it is only incidental to MOS transistor operation, so it seems strange to hang the key visual distinction between the *n*MOS and the *p*MOS symbols on this particular feature of the device. Also, as we argued with the source arrows, the bulk arrows are not really distinct enough for us to be able to tell at a single glance which transistors in a schematic are *n*MOS and which are *p*MOS. In fact, the bulk arrows actually provide less of a distinction than do the source arrows—at least the *n*MOS and *p*MOS transistor symbols with the source arrows are respectively reminiscent of *npn* and *pnp* bipolar transistor symbols, which is probably why the symbols with the source arrows are used by many analog designers who began their careers designing bipolar circuits.

4.10 Incremental MOS Transistor Characteristics

For many circuits that we will study, we shall be interested in investigating the *incremental* behavior of a circuit about some quiescent operating point—that is, how the circuit responds if we change one or more inputs to a circuit by a sufficiently small amount that the circuit acts as a linear system. In effect, the changes have to be small enough that we can replace the nonlinear current–voltage characteristics of each device in the circuit with a multidimensional Taylor series expansion about the circuit's operating point, which we truncate after the first-order (i.e., linear) terms, without making an unacceptably large error in the analysis. For this reason, *incremental analysis* is also called *small-signal analysis.* In this section, we shall

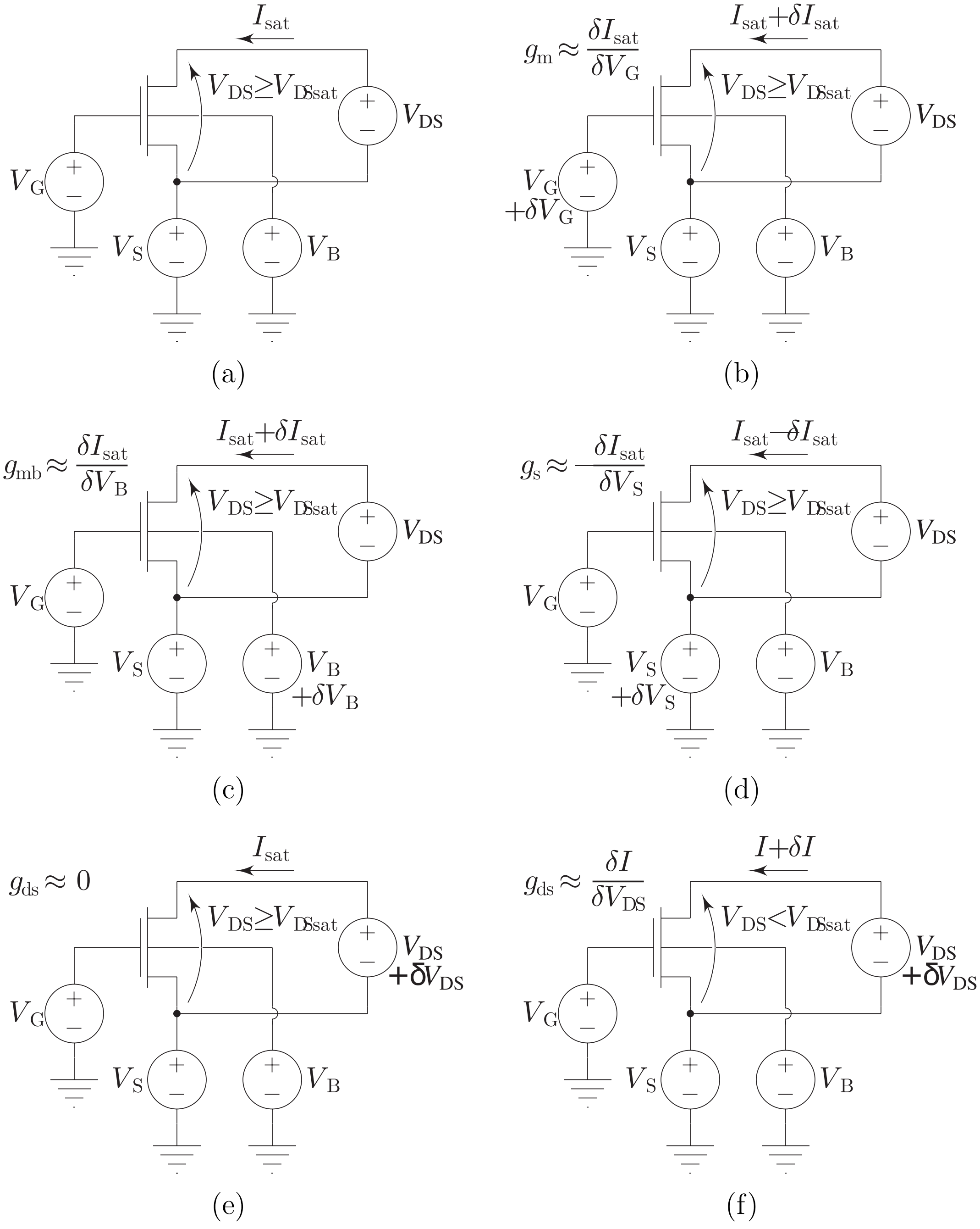

I_{sat}
$V_{\text{DS}} \geq V_{\text{DSsat}}$
V_{DS}
V_{G}
V_{S}
V_{B}
(a)
$g_{\text{m}} \approx \frac{\delta I_{\text{sat}}}{\delta V_{\text{G}}}$
$I_{\text{sat}}+\delta I_{\text{sat}}$
$V_{\text{DS}} \geq V_{\text{DSsat}}$
V_{DS}
V_{G} $+\delta V_{\text{G}}$
V_{S}
V_{B}
(b)
$g_{\text{mb}} \approx \frac{\delta I_{\text{sat}}}{\delta V_{\text{B}}}$
$I_{\text{sat}}+\delta I_{\text{sat}}$
$V_{\text{DS}} \geq V_{\text{DSsat}}$
V_{DS}
V_{G}
V_{S}
V_{B} $+\delta V_{\text{B}}$
(c)
$g_{\text{s}} \approx -\frac{\delta I_{\text{sat}}}{\delta V_{\text{S}}}$
$I_{\text{sat}}-\delta I_{\text{sat}}$
$V_{\text{DS}} \geq V_{\text{DSsat}}$
V_{DS}
V_{G}
V_{S} $+\delta V_{\text{S}}$
V_{B}
(d)
$g_{\text{ds}} \approx 0$
I_{sat}
$V_{\text{DS}} \geq V_{\text{DSsat}}$
V_{DS} $+\delta V_{\text{DS}}$
V_{G}
V_{S}
V_{B}
(e)
$g_{\text{ds}} \approx \frac{\delta I}{\delta V_{\text{DS}}}$
$I+\delta I$
$V_{\text{DS}} < V_{\text{DSsat}}$
V_{DS} $+\delta V_{\text{DS}}$
V_{G}
V_{S}
V_{B}
(f)

Figure 7: Incremental MOS transistor characteristics.

develop expressions for several important *small-signal parameters* for the MOS transistor biased in the saturation region that are valid for all levels of inversion by differentiating the EKV model with respect to various terminal voltages. There are two reasons for focusing on the saturation region. First, many analog circuits are designed so that the MOS transistors are biased into saturation. Second, for those circuits that contain MOS transistors biased into the ohmic region, by exploit the source/drain symmetry of the MOS transistor, source splitting, and superposition, we shall be able to use the incremental relationships that we derive in this section for saturated MOS transistors to analyze circuits containing MOS transistors operating in the ohmic region.

Figure 7a shows what we shall consider as our quiescent operating point for the purposes of computing the incremental properties of an nMOS transistor. If the bulk is connected to a voltage, V_{B}, other than ground, we can express the saturation current approximately by

$$\begin{aligned} I_{\mathrm{sat}} &= I_{\mathrm{s}} \log^2\left(1 + e^{(\kappa(V_{\mathrm{GB}}-V_{\mathrm{T0}})-V_{\mathrm{SB}})/2U_{\mathrm{T}}}\right) \\ &= I_{\mathrm{s}} \log^2\left(1 + e^{(\kappa V_{\mathrm{G}}+(1-\kappa)V_{\mathrm{B}}-\kappa V_{\mathrm{T0}}-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right), \end{aligned}$$

which implies that

$$\log\left(1 + e^{(\kappa V_{\mathrm{G}}+(1-\kappa)V_{\mathrm{B}}-\kappa V_{\mathrm{T0}}-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) = \sqrt{\frac{I_{\mathrm{sat}}}{I_{\mathrm{s}}}} \tag{27}$$

and that

$$e^{(\kappa V_{\mathrm{G}}+(1-\kappa)V_{\mathrm{B}}-\kappa V_{\mathrm{T0}}-V_{\mathrm{S}})/2U_{\mathrm{T}}} = e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}} - 1. \tag{28}$$

From this baseline, we shall consider the effects of a small change in the gate, bulk, source, and drain-to-source voltages in turn.

A *transconductance gain* is a measure of by how much a current somewhere in a circuit changes in response to a change in a voltage at some other location in the circuit. If we increase the gate voltage of the nMOS transistor, the channel current increases, as shown in Fig. 7b. The amount by which the channel current increases in response to a small change in the gate voltage is called the *incremental* or *small-signal* transconductance (or sometimes just simply the transconductance) of the nMOS transistor, which we shall denote by g_{m}. We can obtain an expression for the transconductance of the nMOS transistor in saturation by differentiating the saturation current with respect to the gate voltage. Doing so, we obtain

$$\begin{aligned} g_{\mathrm{m}} &= \frac{\partial I_{\mathrm{sat}}}{\partial V_{\mathrm{G}}} \\ &= I_{\mathrm{s}} \cdot \frac{\partial}{\partial V_{\mathrm{G}}} \log^2\left(1 + e^{(\kappa V_{\mathrm{G}}+(1-\kappa)V_{\mathrm{B}}-\kappa V_{\mathrm{T0}}-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\ &= I_{\mathrm{s}} \cdot 2\log\left(1 + e^{(\kappa V_{\mathrm{G}}+(1-\kappa)V_{\mathrm{B}}-\kappa V_{\mathrm{T0}}-V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\ &\qquad \cdot \frac{e^{(\kappa V_{\mathrm{G}}+(1-\kappa)V_{\mathrm{B}}-\kappa V_{\mathrm{T0}}-V_{\mathrm{S}})/2U_{\mathrm{T}}}}{1 + e^{(\kappa V_{\mathrm{G}}+(1-\kappa)V_{\mathrm{B}}-\kappa V_{\mathrm{T0}}-V_{\mathrm{S}})/2U_{\mathrm{T}}}} \cdot \frac{\kappa}{2U_{\mathrm{T}}}. \end{aligned}$$

We can express the transconductance in terms of the saturation current, I_{sat}, by substituting Eq. 27 and Eq. 28 into Eq. 29 to obtain

$$\begin{aligned} g_{\mathrm{m}} &= \frac{\kappa}{U_{\mathrm{T}}} \cdot I_{\mathrm{s}} \cdot \sqrt{\frac{I_{\mathrm{sat}}}{I_{\mathrm{s}}}} \cdot \frac{e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}} - 1}{e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}} \\ &= \frac{\kappa}{U_{\mathrm{T}}} \cdot \sqrt{I_{\mathrm{s}} I_{\mathrm{sat}}} \cdot \left(1 - e^{-\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}\right) \qquad (29) \\ &\approx \begin{cases} \dfrac{\kappa}{U_{\mathrm{T}}} \cdot I_{\mathrm{sat}}, & I_{\mathrm{sat}} \ll I_{\mathrm{s}} \\ \dfrac{\kappa}{U_{\mathrm{T}}} \cdot \sqrt{I_{\mathrm{s}} I_{\mathrm{sat}}}, & I_{\mathrm{sat}} \gg I_{\mathrm{s}}. \end{cases} \end{aligned}$$

If we were to increase the bulk voltage of the nMOS transistor, the saturation current increases, as shown in Fig. 7c, in much the same way as it did for an increase in the gate voltage. Thus, the bulk also has a transconductance gain, which we shall denote by g_{mb}. By performing a nearly identical sequence of steps to those that we followed in computing g_{m}, we can obtain an expression for the bulk transconductance of the nMOS transistor as

$$\begin{aligned}
g_{\mathrm{mb}} &= \frac{\partial I_{\mathrm{sat}}}{\partial V_{\mathrm{B}}} \\
&= I_{\mathrm{s}} \cdot \frac{\partial}{\partial V_{\mathrm{B}}} \log^2\left(1 + e^{(\kappa V_{\mathrm{G}} + (1-\kappa)V_{\mathrm{B}} - \kappa V_{\mathrm{T0}} - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\
&= \frac{1-\kappa}{U_{\mathrm{T}}} \cdot \sqrt{I_{\mathrm{s}} I_{\mathrm{sat}}} \cdot \left(1 - e^{-\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}\right) \qquad (30) \\
&\approx \begin{cases} \dfrac{1-\kappa}{U_{\mathrm{T}}} \cdot I_{\mathrm{sat}}, & I_{\mathrm{sat}} \ll I_{\mathrm{s}} \\ \dfrac{1-\kappa}{U_{\mathrm{T}}} \cdot \sqrt{I_{\mathrm{s}} I_{\mathrm{sat}}}, & I_{\mathrm{sat}} \gg I_{\mathrm{s}}. \end{cases}
\end{aligned}$$

If we increase the source voltage of an nMOS transistor, the saturation current *decreases*, as shown in Fig. 7d. The *incremental (driving-point) conductance* of the source terminal of the nMOS transistor in saturation is defined a measure of the amount by which the current flowing *into* the source increases in response to a small increase in the source voltage. This quantity is formally given by

$$g_{\mathrm{s}} = \frac{\partial I_{\mathrm{S}}}{\partial V_{\mathrm{S}}},$$

where I_{S} is the current flowing into the source terminal. The saturation current actually flows out of the source terminal and so has the opposite sign from the source current, I_{S}. Thus, we can write the incremental conductance looking into the source of the nMOS transistor as

$$\begin{aligned}
g_{\mathrm{s}} &= \frac{\partial}{\partial V_{\mathrm{S}}}\left(-I_{\mathrm{sat}}\right) \\
&= -\frac{\partial I_{\mathrm{sat}}}{\partial V_{\mathrm{S}}} \\
&= -I_{\mathrm{s}} \cdot \frac{\partial}{\partial V_{\mathrm{S}}} \log^2\left(1 + e^{(\kappa V_{\mathrm{G}} + (1-\kappa)V_{\mathrm{B}} - \kappa V_{\mathrm{T0}} - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\
&= -I_{\mathrm{s}} \cdot 2\log\left(1 + e^{(\kappa V_{\mathrm{G}} + (1-\kappa)V_{\mathrm{B}} - \kappa V_{\mathrm{T0}} - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\
&\qquad \cdot \frac{e^{(\kappa V_{\mathrm{G}} + (1-\kappa)V_{\mathrm{B}} - \kappa V_{\mathrm{T0}} - V_{\mathrm{S}})/2U_{\mathrm{T}}}}{1 + e^{(\kappa V_{\mathrm{G}} + (1-\kappa)V_{\mathrm{B}} - \kappa V_{\mathrm{T0}} - V_{\mathrm{S}})/2U_{\mathrm{T}}}} \cdot \left(-\frac{1}{2U_{\mathrm{T}}}\right) \\
&= \frac{1}{U_{\mathrm{T}}} \cdot I_{\mathrm{s}} \cdot \sqrt{\frac{I_{\mathrm{sat}}}{I_{\mathrm{s}}}} \cdot \frac{e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}} - 1}{e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}} \\
&= \frac{1}{U_{\mathrm{T}}} \cdot \sqrt{I_{\mathrm{s}} I_{\mathrm{sat}}} \cdot \left(1 - e^{-\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}\right) \qquad (31) \\
&\approx \begin{cases} \dfrac{I_{\mathrm{sat}}}{U_{\mathrm{T}}}, & I_{\mathrm{sat}} \ll I_{\mathrm{s}} \\ \dfrac{1}{U_{\mathrm{T}}} \cdot \sqrt{I_{\mathrm{s}} I_{\mathrm{sat}}}, & I_{\mathrm{sat}} \gg I_{\mathrm{s}}. \end{cases}
\end{aligned}$$

From Eq. 29, Eq. 30, and Eq. 31, we can see that very simple relationships hold among g_{m}, g_{mb}, and g_{s} at all levels of inversion. These relationships are given by

$$g_{\mathrm{m}} = \kappa g_{\mathrm{s}}, \quad g_{\mathrm{mb}} = (1-\kappa)\, g_{\mathrm{s}}, \quad \text{and} \quad g_{\mathrm{s}} = g_{\mathrm{m}} + g_{\mathrm{mb}}.$$

In the saturation region, the channel current is approximately independent of the drain-to-source voltage, so we would expect that the incremental conductance between the source and drain is approximately zero, as indicated in Fig. 7e. However, in the ohmic region, the channel current does depend significantly on the drain-to-source voltage, as indicated in Fig. 7f. We can make explicit the dependence of the channel current on V_{DS} in the ohmic region by noting that

$$\begin{aligned}
I &= I_{\mathrm{s}}\left(\log^2\left(1+e^{(\kappa(V_{\mathrm{GB}}-V_{\mathrm{T0}})-V_{\mathrm{SB}})/2U_{\mathrm{T}}}\right) - \log^2\left(1+e^{(\kappa(V_{\mathrm{GB}}-V_{\mathrm{T0}})-V_{\mathrm{DB}})/2U_{\mathrm{T}}}\right)\right)\\
&= I_{\mathrm{sat}} - I_{\mathrm{s}}\log^2\left(1+e^{(\kappa(V_{\mathrm{GB}}-V_{\mathrm{T0}})-V_{\mathrm{SB}}+V_{\mathrm{SB}}-V_{\mathrm{DB}})/2U_{\mathrm{T}}}\right)\\
&= I_{\mathrm{sat}} - I_{\mathrm{s}}\log^2\left(1+e^{(\kappa(V_{\mathrm{GB}}-V_{\mathrm{T0}})-V_{\mathrm{SB}}-V_{\mathrm{DS}})/2U_{\mathrm{T}}}\right)\\
&= I_{\mathrm{sat}} - I_{\mathrm{s}}\log^2\left(1+e^{(\kappa(V_{\mathrm{GB}}-V_{\mathrm{T0}})-V_{\mathrm{SB}})/2U_{\mathrm{T}}}e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}\right)\\
&= I_{\mathrm{sat}} - I_{\mathrm{s}}\log^2\left(1+\left(e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1\right)e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}\right).
\end{aligned}$$

By differentiating this expression for the channel current with respect to V_{DS}, we can express the incremental drain-source conductance of the nMOS transistor in the ohmic region as

$$\begin{aligned}
g_{\mathrm{ds}} &= \frac{\partial I}{\partial V_{\mathrm{DS}}}\\
&= -I_{\mathrm{s}}\cdot\frac{\partial}{\partial V_{\mathrm{DS}}}\log^2\left(1+\left(e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1\right)e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}\right)\\
&= -I_{\mathrm{s}}\cdot 2\log\left(1+\left(e^{\sqrt{I_{\mathrm{sat}}/(W/L)I_{\mathrm{s}}}}-1\right)e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}\right)\\
&\qquad\qquad\cdot\frac{\left(e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1\right)e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}}{1+\left(e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1\right)e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}}\cdot\left(-\frac{1}{2U_{\mathrm{T}}}\right)\\
&= \frac{1}{U_{\mathrm{T}}}\cdot I_{\mathrm{s}}\log\left(1+\left(e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1\right)e^{-V_{\mathrm{DS}}/2U_{\mathrm{T}}}\right)\\
&\qquad\qquad\cdot\frac{e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1}{e^{V_{\mathrm{DS}}/2U_{\mathrm{T}}}+e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1}.
\end{aligned}$$

While this expression is quite a bit more formidable than those that we have developed for g_{m}, g_{mb}, and g_{s}, we shall make two observations based upon it. First, as V_{DS} gets larger than a few U_{T}, the argument of the log goes to unity exponentially, making the log factor go to zero quickly with increasing V_{DS}. Moreover, the last factor also goes to zero exponentially under these circumstances. Thus, our assumption that $g_{\mathrm{ds}} \approx 0$ in the saturation region is generally a good one. Second, we shall examine the value of g_{ds} deep in the ohmic region, when $V_{\mathrm{DS}} = 0$. In this case, we have that

$$g_{\mathrm{ds}}\big|_{V_{\mathrm{DS}}=0} = \frac{1}{U_{\mathrm{T}}}\cdot I_{\mathrm{s}}\log\left(1+\left(e^{\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}-1\right)\cdot 1\right)$$

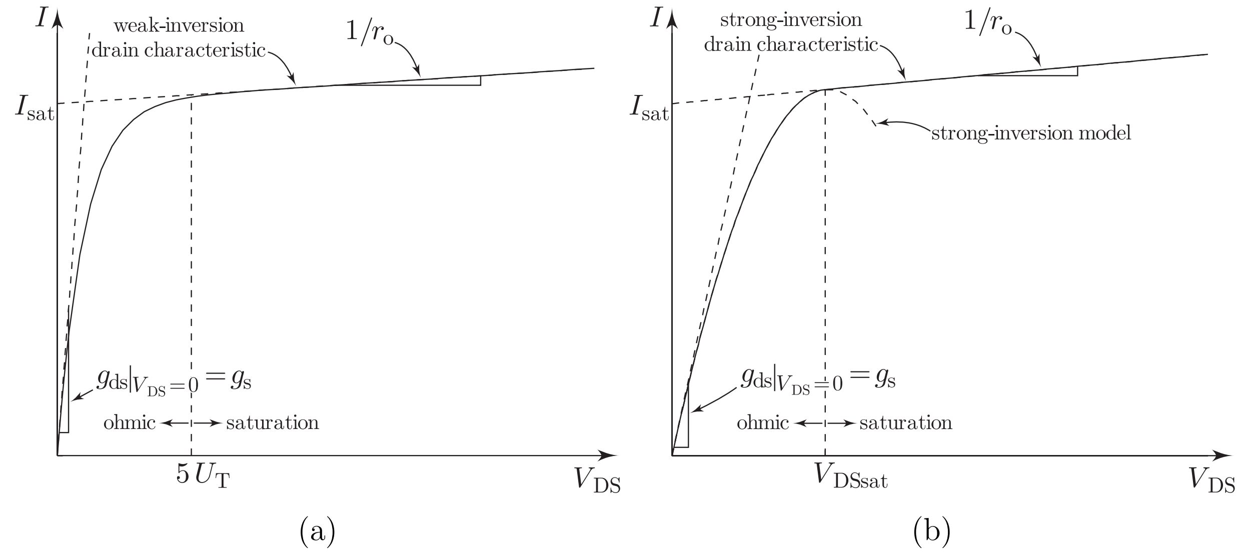


Figure 8: Drain characteristics in (a) weak and (b) strong inversion exhibiting the Early effect.

$$
\begin{aligned}
& \cdot\frac{e^{\sqrt{I_\mathrm{sat}/I_\mathrm{s}}}-1}{1+e^{\sqrt{I_\mathrm{sat}/I_\mathrm{s}}}-1} \\
= \;& \frac{1}{U_\mathrm{T}}\cdot I_\mathrm{s}\log\left(e^{\sqrt{I_\mathrm{sat}/I_\mathrm{s}}}\right)\cdot\frac{e^{\sqrt{I_\mathrm{sat}/I_\mathrm{s}}}-1}{e^{\sqrt{I_\mathrm{sat}/I_\mathrm{s}}}} \\
= \;& \frac{1}{U_\mathrm{T}}\cdot I_\mathrm{s}\cdot\sqrt{\frac{I_\mathrm{sat}}{I_\mathrm{s}}}\cdot\left(1-e^{-\sqrt{I_\mathrm{sat}/I_\mathrm{s}}}\right) \\
= \;& \frac{1}{U_\mathrm{T}}\cdot\sqrt{I_\mathrm{s}I_\mathrm{sat}}\cdot\left(1-e^{-\sqrt{I_\mathrm{sat}/I_\mathrm{s}}}\right) \\
= \;& g_\mathrm{s}.
\end{aligned}
$$

Thus, we have established that the value of g_ds deep in the ohmic region is equal to the incremental conductance of the source terminal in the saturation region. We can make use of this fact to determine the value of the source conductance at any particular saturation current level from a drain characteristic, as indicated in Fig. 8.

4.11 The Early Effect

Drain characteristics of real MOS transistors, especially ones that have lengths that are short relative to the minimum feature size in a given technology, actually exhibit a finite slope in the saturation region at all levels of inversion, as shown in Fig. 8, contrary to the predictions made by the simple EKV model that we have been developing. A markedly similar effect in bipolar transistors was first investigated and modeled by Jim Early in the early 1950s [11], so this effect is commonly named for him. This effect results from an increase in the width of the depletion region surrounding the reverse-biased drain-bulk *p*-*n* junction with an increase in the drain voltage, including in the direction of the channel. The increase in the depletion

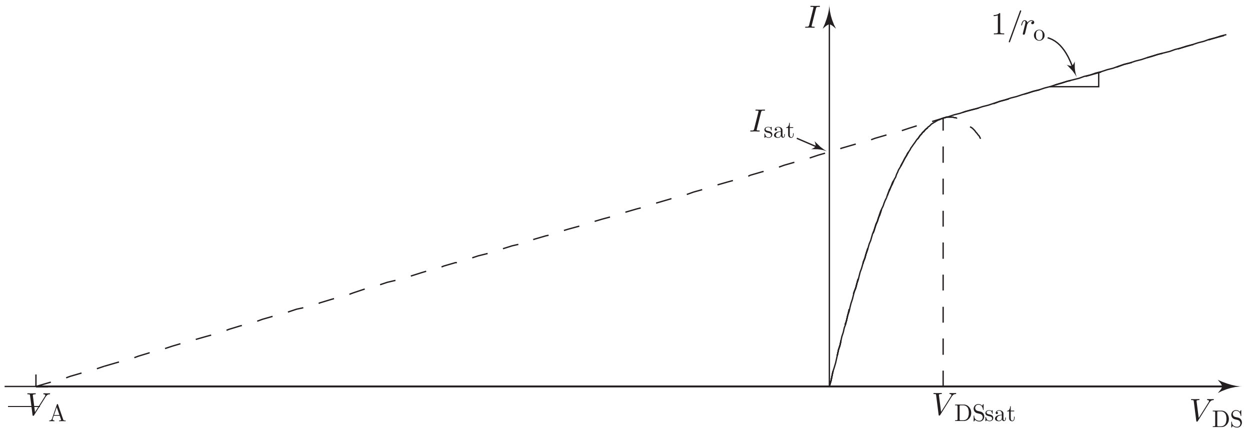


Figure 9: Graphical interpretation of Early's simple model for channel-length modulation.

region width in this direction reduces the effective length of the channel. Because the current in saturation depends inversely on the length of the channel, the net result is an increase in the channel current with increasing drain voltage, and hence drain-to-source voltage. This effect is also commonly referred to as *channel-length modulation*, which is perhaps somewhat more descriptive of its physical origin. Nevertheless, we shall refer to it as the *Early effect.*

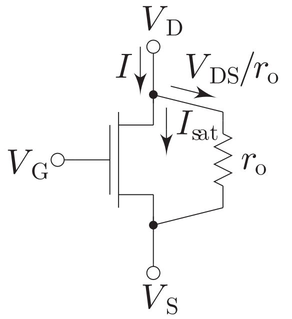
Early proposed a simple empirical expression to model this phenomenon, which translated into language appropriate to the MOS transistor is given by

$$I = I_\mathrm{sat}\left(1 + \frac{V_\mathrm{DS}}{V_\mathrm{A}}\right),$$

where I_sat is the saturation current predicted by the EKV model and V_A is a parameter that has come to be known as the *Early voltage.* A simple graphical interpretation of this model is shown in Fig. 10. The value of V_A is positive by convention, but corresponds to the negative of the V_DS-axis intercept of an extrapolated linear fit to a drain characteristic in the saturation region, as shown in Fig. 10. The larger the value of V_A the less pronounced is the Early effect. The saturation current corresponds to the I-axis intercept of the extrapolated linear fit. On this simple model, the reciprocal of the slope of the current–voltage characteristic in saturation is given by

$$r_\mathrm{o} = \frac{V_\mathrm{A}}{I_\mathrm{sat}}.$$

To the extent that this simple model adequately accounts for the Early effect, we can construct a *large-signal* circuit whose behavior is formally equivalent to Early's empirical expression. We simply connect a linear resistor, which we shall call an *Early-effect resistor*, whose value is given by r_o in parallel with the channel of a saturated MOS transistor, as shown in Fig. 10. We shall assume that the saturated MOS transistor's behavior is predicted by the EKV model, in particular that the incremental conductance looking into its drain terminal is essentially zero. We have accounted for this second-order effect by the explicit addition of a linear parasitic circuit element that is external to an essentially ideal transistor. We have chosen to encapsulate the Early effect in this way so that we can selectively account



$$\begin{aligned} I &= I_{\mathrm{sat}} + \frac{V_{\mathrm{DS}}}{r_{\mathrm{o}}} \\ &= I_{\mathrm{sat}} + V_{\mathrm{DS}} \cdot \frac{I_{\mathrm{sat}}}{V_{\mathrm{A}}} \\ &= I_{\mathrm{sat}} \left(1 + \frac{V_{\mathrm{DS}}}{V_{\mathrm{A}}}\right) \end{aligned}$$

Figure 10: Large-signal equivalent circuit model for the Early effect.

for it in doing a circuit analysis only where it matters. We shall see that neglecting the Early effect when possible *drastically* simplifies the analysis process. While accounting for the Early effect everywhere is not "wrong," it can certainly be counter productive. Unnecessarily accounting for the Early effect greatly complicates the expressions that we obtain from the analysis process, which makes them harder to interpret. It also markedly increases our chances of making an error along the way. Learning to recognize where the Early effect is negligible and where it is not is a *very* valuable bit of circuit intuition to develop.

Given that the postulated physical mechanism underlying the Early effect is an increase in the depletion-region width around the drain-bulk junction and that this increase does not depend on the length of the channel, we should expect that the actual increase in the channel current for a given increase in the depletion-region width will be smaller for longer channel lengths. To see this point more clearly, we write the saturation current using the EKV model for an nMOS transistor as

$$I_{\mathrm{sat}} = I_{\mathrm{s}} \log^2 \left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right),$$

which because of the Early effect becomes

$$\begin{aligned} I'_{\mathrm{sat}} &= \frac{W}{L - \delta L\left(V_{\mathrm{DS}}\right)} I_{\mathrm{s}} \log^2 \left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\ &= \frac{1}{1 - \dfrac{\delta L\left(V_{\mathrm{DS}}\right)}{L}} \cdot I_{\mathrm{s}} \log^2 \left(1 + e^{(\kappa(V_{\mathrm{G}} - V_{\mathrm{T0}}) - V_{\mathrm{S}})/2U_{\mathrm{T}}}\right) \\ &= I_{\mathrm{sat}} \cdot \frac{1}{1 - \dfrac{\delta L\left(V_{\mathrm{DS}}\right)}{L}} \\ &\approx I_{\mathrm{sat}} \left(1 + \frac{\delta L\left(V_{\mathrm{DS}}\right)}{L}\right), \end{aligned}$$

where $\delta L\left(V_{\mathrm{DS}}\right)$ models the functional dependence of the reduction in channel length on V_{DS} and where we have assumed that $\delta L \ll L$, so that we can expand $(1 - x)^{-1}$ in a Taylor series and retain only the linear term. A number of physically plausible empirical forms have been proposed for $\delta L\left(V_{\mathrm{DS}}\right)$ to model solution of the complicated two-dimensional electrostatic

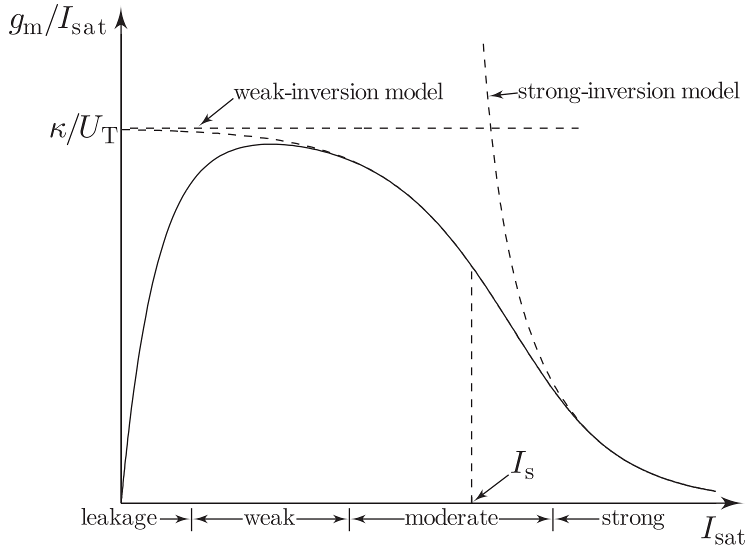


Figure 11: Transconductance generation efficiency for various levels of inversion.

problem that really exists at the drain end of the channel to account for the Early effect. For simplicity, we usually choose to make $\delta L\,(V_{DS}) \propto V_{DS}$, which implies that $V_A \propto L$. We typically find empirically that, for any given current level, the extrapolated Early voltage exhibits a linear dependence on channel length, but usually with a small nonzero offset. It is also very common that we will observe an increase in the effective Early voltage with an increase in the level of inversion, but for hand calculations, we often ignore this dependence and assume that the Early voltage is constant with inversion level.

4.12 Transconductance Generation Efficiency

With the rapid development and proliferation of wireless communication and computing devices, which operate on batteries, low power consumption has steadily risen to the top of the list of key system design specifications. The power efficiency of many circuits is related directly to the ratio of the transconductance gain of the devices in the circuit to the quiescent bias currents flowing through them. The transconductance factors directly into important circuit specifications, such as gain and bandwidth, while the quiescent bias current factors directly into power consumption. This ratio is often called *transconductance per unit current*. It is also termed *transconductance generation efficiency*, because it measures how efficiently a transistor converts current into transconductance. An important principle involved in designing power efficient circuits is to use devices with high transconductance generation efficiency or to bias devices where the transconductance generation efficiency is as large as possible [12–14].

The bipolar transistors, whose transconductance gain is given by $g_m = I_C/U_T$, where I_C is the quiescent collector current and U_T is the thermal voltage, has the highest transconductance generation efficiency, of any active device—its value is a constant, given by $g_m/I_C = 1/U_T$. For this reason, bipolar transistors are still generally preferred to MOS transistors in the design of power efficient wireless communication circuits. For MOS transistors, transconductance generation efficiency is a strong function of bias level, as shown in Fig. 11. Its value is maximum in weak inversion, given by κ/U_{T}, and begins to decrease monotonically as the device enters moderate inversion and then strong inversion. For short-channel MOS transistors that exhibit velocity saturation in strong inversion, the transconductance generation efficiency degrades even faster than is shown in Fig. 11. In attempting to design circuits that are power efficient, we should bias transistors where their transconductance generation efficiency is highest, which suggests that we bias MOS transistors in either weak or moderate inversion, rather than in strong inversion.

By using the expression that we derived in Section 4.10 for the transconductance gain of the MOS transistor, we can develop a simple analytical expression for the transconductance generation efficiency of an MOS transistor in terms of its saturation current that is valid for all levels of inversion. To do so, we simply divide Eq. 29 by I_{sat} to obtain

$$\begin{aligned}\frac{g_{\mathrm{m}}}{I_{\mathrm{sat}}} &= \frac{\kappa}{U_{\mathrm{T}}}\cdot\sqrt{\frac{I_{\mathrm{s}}}{I_{\mathrm{sat}}}}\cdot\left(1-e^{-\sqrt{I_{\mathrm{sat}}/I_{\mathrm{s}}}}\right)\\ &\approx \begin{cases}\dfrac{\kappa}{U_{\mathrm{T}}}, & I_{\mathrm{sat}} \ll I_{\mathrm{s}}\\ \dfrac{\kappa}{U_{\mathrm{T}}}\cdot\sqrt{\dfrac{I_{\mathrm{s}}}{I_{\mathrm{sat}}}}, & I_{\mathrm{sat}} \gg I_{\mathrm{s}}.\end{cases}\end{aligned}$$

It is possible to use this expression to choose systematically W/L ratios for MOS transistors in analog circuits in order to achieve a desired degree of power efficiency [13].

References

[1] J. E. Meyer, "MOS Models and Circuit Simulation," *RCA Review*, vol. 32, no. 1, pp. 42–63, 1971.

[2] M. A. Maher and C. A. Mead, "A Physical Charge-Controlled Model for MOS Transistors," in *ARVLSI: Proceedings of the 1987 Stanford Conference*, P. Losleben, Ed., pp. 211–229. MIT Press, Cambridge, MA, 1987.

[3] M. A. Maher, *A Charge-Controlled Model for MOS Transistors*, Ph.D. thesis, Caltech, Pasadena, CA, 1989.

[4] C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, MA, 1989.

[5] E. A. Vittoz, "Micropower Techniques," in *Design of Analog-Digital Circuits for Telecommunications and Signal Processing*, J. E. Franca and Y. Tsividis, Eds., pp. 53–96. Prentice-Hall, Englewood Cliffs, NJ, 1994.

[6] K. Bult, "Basic CMOS Circuit Techniques," in *Analog VLSI Signal and Information Processing*, M. Ismail and T. Feiz, Eds., pp. 12–56. McGraw-Hill, New York, 1994.

[7] H. Wallinga and K. Bult, "Design and Analysis of CMOS Signal Processing Circuits by Means of a Graphical MOST Model," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 3, pp. 672–680, 1989.

[8] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, no. 1, pp. 83–114, 1995.

[9] C. G. Montoro and A. I. A. Cunha, "A Current-Based MOSFET Model for Integrated Circuit Design," in *Low-Voltage/Low-Power Integrated Circuits and Systems*, E. Sánchez-Sinencio and A. G. Andreou, Eds., pp. 7–55. IEEE Press, Piscataway, NJ, 1999.

[10] K. Bult and G. J. G. M. Geelen, "An Inherently Linear and Compact MOST-Only Current Division Technique," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 6, pp. 1730–1735, 1992.

[11] J. M. Early, "Effects of Space-Charge Layer Widening in Junction Transistors," *Proceedings of the IRE*, vol. 42, no. ??, pp. 1401–??, 1952.

[12] A. G. Andreou and K. A. Boahen, "Neural Information Processing II," in *Analog VLSI Signal and Information Processing*, M. Ismail and T. Feiz, Eds., pp. 358–413. McGraw-Hill, New York, 1994.

[13] F. Silveira, D. Flandre, and P. G. A. Jespers, "A gm/ID Based Methodology for the Design of CMOS Analog Circuits and Its Application to the Synthesis of a Silicon-on-Insulator Micropower OTA," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 9, pp. 1314–1319, 1996.

[14] A. G. Andreou, "Exploiting Device Physics in Circuit Design for Efficient Computational Functions in Analog VLSI," in *Low-Voltage/Low-Power Integrated Circuits and Systems: Low-Voltage Mixed-Signal Circuits*, E. Sánchez-Sinencio and A. G. Andreou, Eds., pp. 85–132. IEEE Press, Piscataway, NJ, 1999.