Chapter 17: Short-Channel Effects and Device Models

17.1 Scaling Theory

17.2 Short-Channel Effects

17.3 MOS Device Models

Background

- The square-law characteristics derived for MOSFETs provide moderate accuracies for devices with minimum channel lengths of greater than several microns.

- As device dimensions continue to scale down, reaching below 12 nm, higher order effects necessitate more complex models.

- Our objective here is to provide a basic understanding of short-channel effects and review some of the SPICE models developed to reflect such phenomena.

- We first describe the ideal scaling theory of MOS transistors. Next, we study short-channel effects.

Scaling Theory

- The two principal reasons for the dominance of CMOS technology in today’s semiconductor industry are the zero static power dissipation of CMOS logic and the scalability of MOSFETs.

- The ideal scaling theory follows three rules:

(1) reduce all lateral and vertical dimensions by α (>1)

(2) reduce the threshold voltage and the supply voltage by α

(3) increase all of the doping levels by α

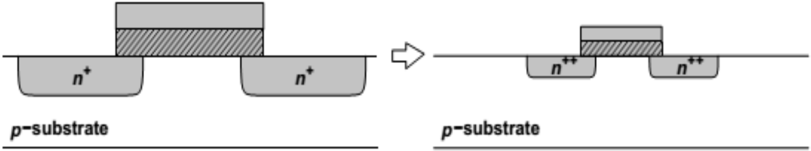


Figure 17.1. Ideal scaling of MOS transistor.

Constant-field scaling

- **Since the dimensions and voltages scale together, all electric fields in the transistor remain constant, hence the name "constant-field scaling."**
- **Note that W, L, t_{ox}, V_{DD}, V_{TH} and the depth and perimeter of the source and drain junctions scale down by α.**
- **Saturation drain current of a square-law device after scaling :**

$$\begin{aligned} I_{D,scaled} &= \frac{1}{2}\mu_n(\alpha C_{ox})\left(\frac{W/\alpha}{L/\alpha}\right)\left(\frac{V_{GS}}{\alpha}-\frac{V_{TH}}{\alpha}\right)^2 \\ &= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS}-V_{TH})^2\frac{1}{\alpha}, \end{aligned}$$

- **We observe that the current capability of the transistor drops by a factor of α.**

Advantages of Scaling

- The advantage of scaling lies in the reduction of capacitances and power dissipation.
- The total channel capacitance is :

$$\begin{aligned} C_{ch,scaled} &= \frac{W}{\alpha}\frac{L}{\alpha}(\alpha C_{ox}) \\ &= \frac{1}{\alpha} W L C_{ox}. \end{aligned}$$

- To calculate the source/drain junction capacitance, we first analyze the effect of ideal scaling on the total width of the depletion region.
- Recall that this width is given by :

$$W_d = \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A}+\frac{1}{N_D}\right)(\phi_B + V_R)},$$

- Where N_A and N_D denote the doping levels of the two sides of the junction.

Depletion region capacitance

- V_R is the reverse-bias volta

$$\phi_B = V_T \ln(N_A N_D/n_i^2)$$

- Th ial, Φ_B is fa $N_A N_D$ is

$$W_{d,scaled} \approx \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{\alpha N_A}+\frac{1}{\alpha N_D}\right)\frac{V_R}{\alpha}}$$

$$\approx \frac{1}{\alpha}\sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A}+\frac{1}{N_D}\right)V_R}.$$

- For now, we assume V_R >>

S/D junction capacitances

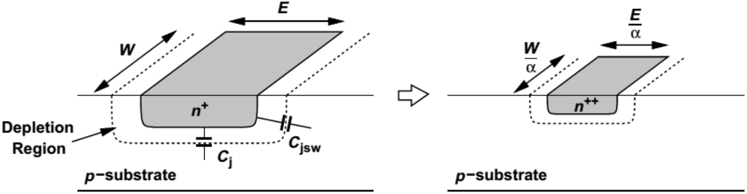


Figure 17.2. Scaling of S/D junction capacitances.

- **The bottom-plate capacitance of the S/D junction (per unit area), C_j, increases by a factor of α.**
- **The sidewall capacitance(per unit width), C_{jsw} remains constant, as the depth of the junction is reduced by α.**

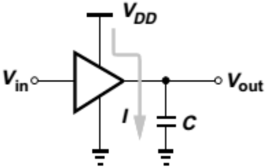
$$\begin{aligned} C_{S/D,scaled} &= \frac{W}{\alpha}\frac{E}{\alpha}(\alpha C_j) + 2\left(\frac{W}{\alpha} + \frac{E}{\alpha}\right)(C_{jsw}) \\ &= [WEC_j + 2(W+E)C_{jsw}]\frac{1}{\alpha}. \end{aligned}$$

- **All of the capacitances therefore decrease by the scaling factor.**

Delay

- Approximating the delay of a (

$$T_{d,scaled} = \frac{C/\alpha}{I/\alpha}\frac{V_{DD}}{\alpha} = \left(\frac{C}{I}V_{DD}\right)\frac{1}{\alpha}.$$



CMOS inverter.

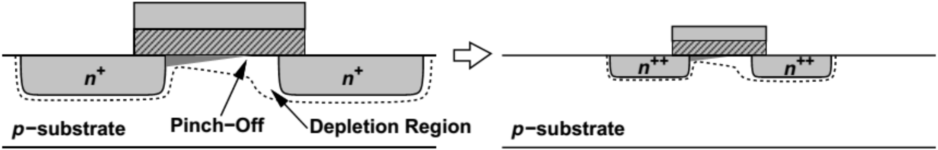
$$P_{scaled} = f(C/\alpha)(V_{DD}/\alpha)^2 = fCV_{DD}^2/\alpha^3$$

- We conclude that the speed of

Effect of ideal scaling

- Effect of ideal scaling:- Writing

$$
\begin{aligned}
g_{m,scaled} &= \mu(\alpha C_{ox})\frac{W/\alpha}{L/\alpha}\frac{V_{GS}-V_{TH}}{\alpha}\\
&= \mu C_{ox}\frac{W}{L}(V_{GS}-V_{TH}),
\end{aligned}
$$



Effect of ideal scaling

- Since $\lambda = (\Delta L/L)/V_{DS}$, λ inc

$$r_{O,scaled} = \frac{1}{\alpha\lambda\frac{I_D}{\alpha}} = \frac{1}{\lambda I_D}.$$

- Thus, the intrinsic gain, $g_m r_o$

$g_m r_o$ has dropped considerabl

The greatest impact of scaling

Short-Channel Effects

- In order to appreciate the need for sophisticated device models, we briefly study some of the phenomena that manifest themselves for short channels

- Small-geometry effects arise because five factors deviate the scaling from the ideal scenario:

(1) The electric fields tend to increase because the supply voltage has not scaled proportionally.

(2) The built-in potential term Φ_B is neither scalable nor negligible.

(3) The depth of S/D junctions cannot be reduced easily.

(4) The mobility decreases as the substrate doping increases.

(5) The subthreshold slope (described below) is not scalable.

Threshold Voltage Variation

- The choice of the threshold voltage is based on the device performance in typical circuit applications.

- The upper bound is roughly equal to $V_{DD}/4$ to avoid degrading the speed of digital CMOS gates.

- The lower bound is determined by several factors:

  (1) the subthreshold behavior.

  (2) variation with temperature and process.

  (3)dependence upon the channel length.

- Let us first consider the subthreshold behavior.

Subthreshold behavior

- For long-channel devices, the subthreshold drain current can be expressed as

$$I_D = \mu C_d \frac{W}{L} V_T^2 \left(\exp \frac{V_{GS} - V_{TH}}{\zeta V_T} \right) \left(1 - \exp \frac{-V_{DS}}{V_T} \right)$$

where $C_d = \sqrt{\epsilon_{si} q N_{sub} / (4\phi_B)}$, denotes the capacitance of the depletion region under the gate area, $V_T = kT/q$, and $\zeta = 1 + C_d / C_{ox}$.

- This equation reveals two interesting properties :
- First, as V_{DS} exceeds a few V_T , I_D becomes independent of the drain-source voltage.
- Second, under this condition the slope of I_D on a logarithmic scale equals

$$\frac{\partial(\log_{10} I_D)}{\partial V_{GS}} = (\log_{10} e) \frac{1}{\zeta V_T}$$

Subthreshold behavior

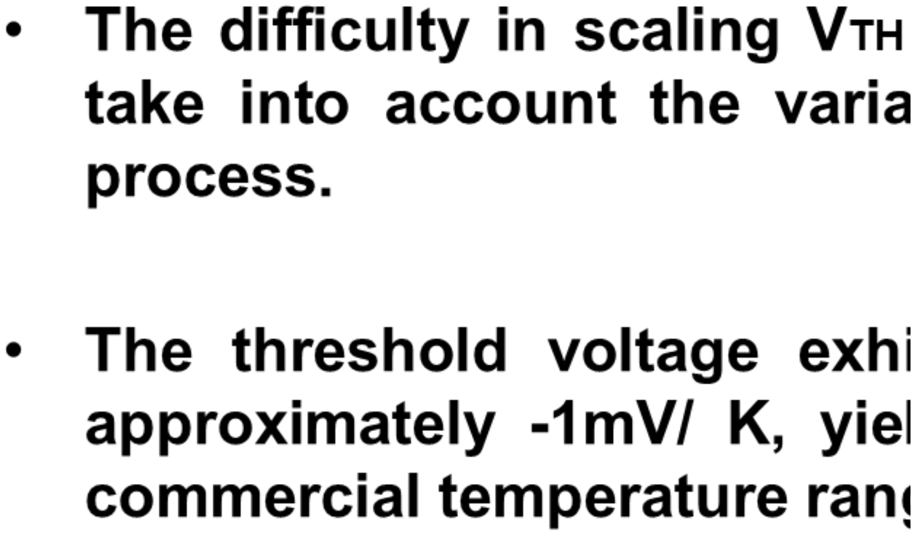
- The inverse of this quantity (slope of I_D on a logarithmic scale) is usually called the "subthreshold slope", *S*:

$$S = 2.3V_T \left(1 + \frac{C_d}{C_{ox}}\right) \ \text{V/dec}$$

- In order to turn off the transistor by lowering V_{GS} below V_{TH}, *S* must be as small as possible, i.e., *Cd / Cox* must be minimized.

- The relatively constant magnitude of *S* severely limits the scaling of the threshold voltage.
- For example : A subthreshold slope of 80 mV/dec imposes a lower bound of 400 mV for V_{TH} if the "off current" must be roughly five orders of magnitude lower than the "on current."

Subthreshold behavior

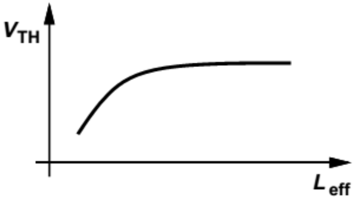


- The difficulty in scaling VTH take into account the varia process.

- The threshold voltage exhi approximately -1mV/ K, yiel commercial temperature ran

Effect of channel length on threshold

- An interesting phenomenon observed in scaled transistors is the dependence of the threshold voltage on the channel length.

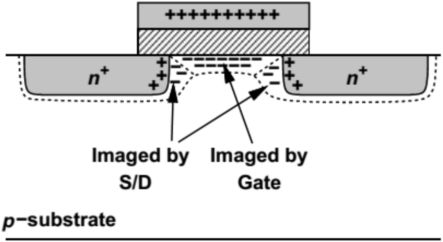


Variation of threshold with channel length.

- As shown in the figure, transistors fabricated on the same wafer but with different lengths yield lower V_{TH} as L decreases.

- This is because the depletion regions associated with the source and drain junctions protrude into the channel area considerably, thereby reducing the immobile charge that must be imaged by the charge on the gate

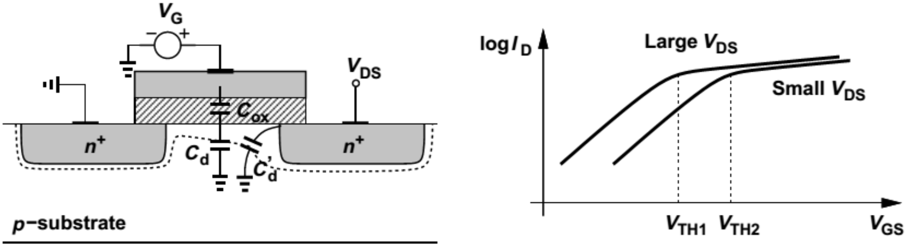
Effect of channel length on threshold



Charge sharing between source/drain depletion regions and the channel depletion region.

- **Part of the immobile charge in the substrate is now imaged by the charge inside the source and drain areas rather than by the charge on the gate.**
- **As a result, the gate voltage required to create an inversion layer decreases**
- **Since the channel length cannot be controlled accurately during fabrication, this effect introduces additional variations in V_{TH}.**

Drain-induced barrier lowering (DIBL)

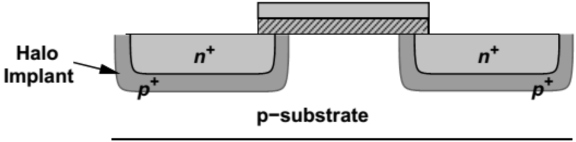


(a) (b)

(a) DIBL in a short-channel device, (b) effect of DIBL on current characteristic.

- In weak inversion, as the ga potential becomes more pos source region.
- In essence, the drain introd

Reverse Short-Channel Effect



MOS structure with halo implant.

- In nanometer CMOS technologies, the threshold voltage decreases as the channel length increases from its minimum value.
- Let us consider the cross section of a modern device, shown in the figure, wherein a "halo" implant of heavy doping surrounds the source and drain junctions.
- This implant reduces the penetration of the drain depletion region into the channel area, thereby improving the device characteristics.

Reverse Short-Channel Effect

- **The threshold voltage is a function of the substrate doping level, *Nsub*. We have**

$$V_{TH} = \phi_{MS} + 2\phi_F + \frac{Q_{dep}}{C_{ox}}$$

where both $\phi_F = (kT/q)\ln(N_{sub}/n_i)$ $Q_{dep} = \sqrt{4q\epsilon_{si}|\phi_F|N_{sub}}$ **increase as *Nsub* increases.**

- **Due to the nonuniform substrate doping along the channel the "local" threshold voltage also varies from the source to the drain.**

- **We can take the average along the channel to obtain an overall threshold for a given device structure**

Mobility Degradation

- At large gate-source vol
developed between the gat
charge carriers to a narrow
interface -

leading to more carrier scatt

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})}$$

- Since scaling has sub

Mobility Degradation

- In addition to lowering the current capability and transconductance of MOSFETs, mobility degradation deviates the I/V characteristic from the simple square-law behavior.

$$I_D = \frac{1}{2}\frac{\mu_0 C_{ox}}{1+\theta(V_{GS}-V_{TH})}\frac{W}{L}(V_{GS}-V_{TH})^2$$

- And assuming $\theta(V_{GS}-V_{TH}) \ll 1$, we obtain

$$\begin{aligned} I_D &\approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[1-\theta(V_{GS}-V_{TH})](V_{GS}-V_{TH})^2 \\ &\approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[(V_{GS}-V_{TH})^2-\theta(V_{GS}-V_{TH})^3]. \end{aligned}$$

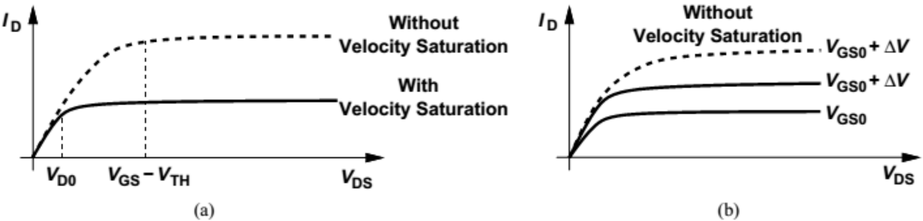
- This is a rough approximation but it reveals the existence of higher harmonics in the drain current.
- The mobility degradation with the vertical field affects the device transconductance as well

Effect of Velocity Saturation

An important consequence is that, as V_{GS} increases, the drain current saturates well before pinch-off occurs.

As shown in the figure carriers reach velocity saturation if V_{DS} exceeds $V_{D0} < V_{GS} - V_{TH}$, yielding a constant current quite lower than that obtained if the device saturated for $V_{DS} > V_{GS} - V_{TH}$.

Since an increment in V_{GS} gives a smaller increment for I_D when velocity saturation occurs, the transconductance is also lower than that predicted by the square law



Effect of velocity saturation: (a) premature drain current saturation, (b) reduction of transconductance.

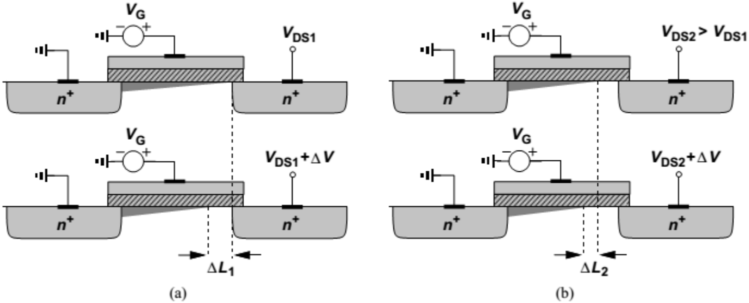
Hot Carrier Effects

- While the average velocity of carriers saturates at high fields, the instantaneous velocity and hence the kinetic energy of the carriers continue to increase, especially as they accelerate towards the drain.
  - These are called “hot” carriers.

- In the vicinity of the drain region, hot carriers may “hit” the silicon atoms at high speeds, thereby creating impact ionization. As a result, new electrons and holes are generated, with the electrons absorbed by the drain and the holes by the substrate. Thus, a finite drain-substrate current appears.
- Also, if the carriers acquire a very high energy, they may be injected into the gate oxide and even flow out the gate terminal, introducing a gate current.

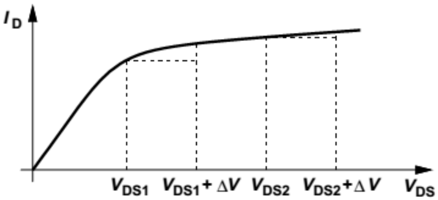
Output Impedance Variation with V_{DS}

As V_{DS} increases and the pinch-off point moves toward the source, the rate at which the depletion region around the source becomes wider decreases, resulting in a higher incremental output impedance.



Decrement in channel length for (a) small V_{DS} and (b) large V_{DS}.

Output Impedance Variation with V_{DS}

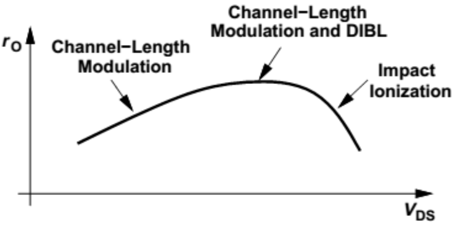


- In this regime, the output impedance can be approximated as

$$r_O = \frac{2L}{1-\dfrac{\Delta L}{L}} \frac{1}{I_D} \sqrt{\frac{qN_B}{2\epsilon_{si}}(V_{DS}-V_{DS,sat})}$$

- Where *$V_{D,sat}$* is the drain-source voltage at the onset of pinch-off.

Output Impedance Variation with V_{DS}



Overall variation of output resistance as a function of V_{DS}

- In In short-channel devices, as VDS increases further, drain-induced barrier lowering becomes significant, reducing the threshold voltage and increasing the drain current.
- This effect roughly cancels that expressed by r_0 equation, giving a relatively constant output impedance.
- At sufficiently high drain voltages, impact ionization near the drain produces a large current (flowing from the drain into the substrate), in essence lowering r_0.

MOS Device Models

- **MOS device modeling continues to pose a challenge—especially for high-frequency operation.**

- **Our objective is to develop a basic understanding of some of the models to the extent necessary for simulations.**

- **The utility of a model is given by the accuracy it provides in various regions of operation for different device dimensions, the ease with which its parameters can be measured, and the efficiency that it allows in simulations.**

Level 1 Model

- **Also known as the Shichman and Hodges Model and is based on the following equations:**

$$I_D = \frac{1}{2}K_P\frac{W}{L-2L_D}[2(V_{GS}-V_{TH})V_{DS}-V_{DS}^2](1+\lambda V_{DS}) \quad \text{Triode Region}$$

$$I_D = \frac{1}{2}K_P\frac{W}{L-2L_D}(V_{GS}-V_{TH})^2(1+\ lambdaV_{DS}) \quad \text{Saturation Region}$$

where $K_P = \mu C_{ox}$ and $V_{TH} = V_{TH0}+\gamma(\sqrt{2\phi_B - V_{BS}}-\sqrt{2\phi_B})$

- **Note that this model does not include subthreshold conduction or any short-channel effects.**

- **Since in the simple model, C$_{GS}$ abruptly changes from (2/3)*WL*C$_{ox}$ + *W*C$_{ov}$ in saturation to (1/2)*WL*C$_{ox}$ + *W*c$_{ov}$ in the triode region most computation algorithms experience convergence difficulties here.**

Level 2 Model

- Since the inversion layer must image the charge o layer vanishes in the dire depletion region must er

$$I_D = \mu C_{ox}\frac{W}{L}\{(V_{GS}-V_{TH0})V_{DS}-\frac{V_{DS}^2}{2} -\frac{2}{3}\gamma[(V_{DS}-V_{BS}+2\phi_F)^{3/2}-(-V_{BS}+2\phi_F)^{3/2}]\}$$

- Performing the integratic threshold voltage yields

Level 2 Model

- **Moreover, for small V_{DS}, the equation reduces to that of the Level 1 model, but for large V_{DS} the drain current is less than that predicted by the square law.**

- **It can also be shown that the edge of the saturation region is given by**

$$V_{D,sat} = V_{GS} - V_{TH0} - \phi_F + \gamma^2 \left[1 - \sqrt{1 + \frac{2}{\gamma^2}(V_{GS} - V_{TH0} + \phi_F)} \right]$$

- **In the saturation region, the drain current is**

$$I_{DS} = I_{D,sat} \frac{1}{1 - \lambda V_{DS}}$$

Where $I_{D,sat}$ is calculated from I_D for $V_{DS} = V_{DS,sat}$.

Level 2 Model

- **Modeling channel-length modulation or, more generally, the finite output impedance has always remained a difficult problem. Representing such phenomena by only λ is far from accurate.**

- **Using simple relationships for the depletion region of a pn junction, we can write**

$$\Delta L = \sqrt{\frac{2\epsilon_{si}}{qN_{sub}}[\phi_B + (V_{DS} - V_{D,sat})]}$$

Where $V_{D,sat}$ denotes the pinch-off voltage.

- **The principal difficulty in the above approach is that both the drain current and its derivative are discontinuous at the edge of the triode region.**

Level 2 Model

- To resolve this issue, ∆L is actually obtained by a "fixed-up" equation:

$$\Delta L = \sqrt{\frac{2\epsilon_{si}}{qN_{sub}}\left(V_1 + \sqrt{1 + V_1^2}\right)}$$

- Where $V_1=(V_{DS} - V_{D,sat})/4$. The channel-length modulation coefficient is expressed as $\lambda = \Delta L/(L\ V_{DS})$
- The Level 2 model also includes the degradation of the mobility with the vertical field in the channel. The mobility is calculated from

$$\mu_s = \mu_0 \left(\frac{\epsilon_{si}}{C_{ox}} \cdot \frac{U_c}{V_{GS} - V_{TH} - U_t V_{DS}}\right)^{U_e}$$

- Where U_c denotes the gate-channel critical electric field, U_t is a fitting parameter between 0 and 0.5, and U_e is an exponent in the vicinity of 0.15.

Level 2 Model

- The subthreshold behavior implemented in the Level 2 model defines a voltage V_{on} as

$$V_{on} = V_{TH} + \zeta V_T, \text{where } \zeta = 1 + (qN_{FS}/C_{ox}) + C_d/C_{ox}$$

and N_{FS} is an empirical constant.

- The drain current is then expressed as:

$$I_{DS} = I_{on} \exp \frac{V_{GS} - V_{on}}{\zeta V_T}$$

Where I_{on} is the drain current calculated in strong inversion for V_{GS}=V_{on}.

Level 3 Model

- The mobility equation in

$$\mu_1 = \frac{\mu_{eff}}{1 + \frac{\mu_{eff} V_{DS}}{v_{max} L_1}}$$

fiel

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})}$$

nd i

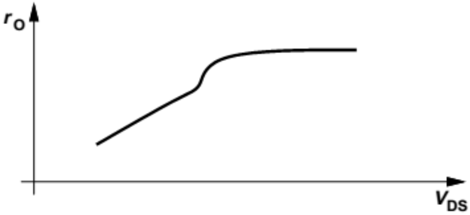
$$I_D = \mu_1 C_{ox} \frac{W_{eff}}{L_{eff}} \left[V_{GS} - V_{TH0} - \left(1 + \frac{F_s \gamma}{4\sqrt{2\phi_F - V_{BS}}} + F_n \right) \frac{V'_{DS}}{2} \right] V'_{DS}$$

and v_{max} denotes the

Level 3 Model

- Level 3 model, as with the Level 2 model, exhibits moderate accuracy for wide, short transistors but suffers from large errors for longer channels.

- An important drawback of the Level 3 model is the discontinuity of the derivative of I_D with respect to V_{DS} at the edge of the triode region, leading to large errors in the calculation of the output impedance.



Kink in output resistance in Level 3 model.

BSIM Series

- The philosophy behind the Level 1-3 models was to express the device behavior by means of equations that originated from the physical operation.
- BSIM adopted a different approach: numerous empirical parameters were added so as to simplify the equations—but at the cost of losing touch with the actual device operation.
- An interesting feature of BSIM is the addition of a simple equation to represent the geometry dependence of many of the device parameters.

BSIM Series

- The g pression

$$P = P_0 + \frac{\alpha_P}{L_{eff}} + \frac{\beta_P}{W_{eff}}$$

$$\mu = \mu_0 + \frac{\alpha_\mu}{L_{eff}} + \frac{\beta_\mu}{W_{eff}}$$

w s the valu wide transistor (P = P_0 if Le fitting factors. For example as:

BSIM Series

(2) the threshold voltage is modified for substrates with nonuniform doping.

(3) the currents in the weak and strong inversion regions are derived such that their values and first derivatives are continuous.

(4) to simplify the drain current equations, new expressions are devised for velocity saturation, dependence of mobility upon the lateral field, and the saturation voltage.

BSIM2

- The next model in the BSIM series is BSIM2. Requiring approximately 70 parameters, this version employs new expressions for mobility, drain current, and subthreshold conduction.

- It also represents the output impedance more accurately by incorporating both channel-length modulation and drain-induced barrier lowering.

- For short, narrow transistors, BSIM2 suffers from large errors in the triode region and even substantial “kinks” in the saturation region.

BSIM3

- The trend in BSIM and BSIM2, namely, expressing the device behavior by means of empirical equations that bear little relation to the physical phenomena, eventually created difficulties in modeling short-channel devices.

- Consequently, the next generation, BSIM3, has returned to the physical principles of device operation while maintaining many of the useful features of BSIM and BSIM2.

- BSIM3 itself has rapidly gone through several versions, requiring approximately 180 parameters in the third one.

- BSIM3 provides reasonable accuracy for subthreshold and strong inversion operation while still suffers from large errors in predicting the output impedance

Other Models

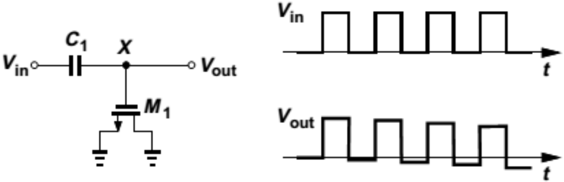
- In addition to the Level 1-3 models and the four generations of BSIM, a number of other MOS models have been introduced.
- Among these, HSPICE Level 28, MOS9, and the EnzKrummenacher-Vittoz (EKV) model are the most notable.
- For example, the HSPICE Level 28 model improves the dependence of accuracy upon device dimensions by expressing the parameters as:

$$P = P_0 + \alpha\left(\frac{1}{L} - \frac{1}{L_{ref}}\right) + \beta\left(\frac{1}{W} - \frac{1}{W_{ref}}\right) + \gamma\left(\frac{1}{L} - \frac{1}{L_{ref}}\right)\left(\frac{1}{W} - \frac{1}{W_{ref}}\right)$$

Where Lref and Wref denote the dimensions of a "reference" device, i.e., a transistor whose characteristics have been measured.

Charge and Capacitance Modeling

- The simple gate capacitance model described for the Level 1 model, called the Meyer capacitance model, suffers from many shortcomings even for long-channel devices.
- In transient SPICE analyses, such a model does not conserve charge (!), thereby introducing errors in the simulation.



Annihilation of charge in simulation.

- A periodic re ge divider consisting of an ideal capacitor and a MOSFET experiences “droop” at the output because in every period some charge at node X is lost.

Charge and Capacitance Modeling

- This voltage droop effect arises from the calculation of charge by integrating capacitor voltages with respect to time, an operation that accumulates small errors in the simulation.
- To minimize this type of error, the simulation algorithm can be modified such that it first computes the charge in the inversion layer and the depletion region and subsequently partitions the charge among the device capacitances.

- Another issue in the Meyer charge model is as follows -
- The assumption that in triode region $C_{GS} = C_{GD} = (1/2)WLC_{ox} + WC_{OV}$, and in saturation region $C_{GS} = (2/3)WLC_{ox} + WC_{OV}$ and $C_{GD} = WC_{OV}$ is inaccurate for short channel devices, requiring flexible partitioning for ease of curve fitting

Temperature Dependence

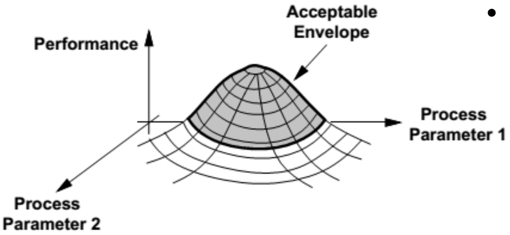
- Many parameters of MOS transistors vary with temperature, making it difficult to maintain a reasonable fit between measured and simulated behavior across a wide temperature range.

- In the Level 1-3 models as well as BSIM and BSIM2, the following parameters have temperature dependence:
- V_{TH}, built-in potential of S/D junctions, the intrinsic carrier concentration of silicon (n_i), the bandgap energy (E_g), and the mobility. Most equations are empirical, e.g.,

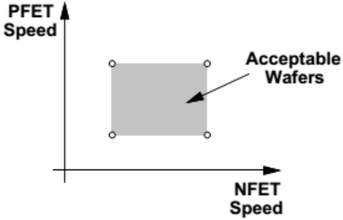
$$E_g = 1.16 - \frac{7.02 \times 10^{-4} T^2}{T + 1108}$$

$$\mu = \mu_0 \left(\frac{300}{T} \right)^{3/2},$$

Process Corners



Performance envelope as a function of process parameters.



Process corners based on speed of NMOS and PMOS devices.

- **Unlike bipolar transistors, MOSFETs suffer from substantial parameter variations from wafer to wafer.**

- **Process engineers guarantee a performance envelope for the devices.**

The idea is to constrain the speed envelope of the NMOS and PMOS transistors to a rectangle defined by four corners – called process corners.