Chapter 1

Subthreshold MOS for Ultra-Low Power

This Chapter provides a brief review on modeling of MOSFET devices especially for weak-inversion (WI) devices.\(^1\) The main issues associated with WI design such as variation due to the PVT, mismatch effects, and device noise are briefly addressed. Meanwhile, a review on the main problems for implementing ULP CMOS circuits is provided. At the end of the Section, an analytical approach for systematic design of digital CMOS circuits operating in WI region with optimum energy consumption and acceptable reliability is proposed.

1.1 MOS Technology

The first proposal for implementing metal-oxide-semiconductor field-effect transistors (MOSFETs) can be traced back to 1930, when Lilienfeld and Heil patented the initial concept independently [1]-[3]. However, successful implementation demonstrated after a few decades in 1960. Simple topology of MOSFETs in addition to their small area, makes it possible to implement very large-scale integrated (VLSI) circuits. This property is especially demanding for implementing digital systems with very powerful processing capabilities. Commercial requirements have pushed the need for fabricating ICs with more powerful processing capabilities or more number of devices per chip area for the past couple of decades as depicted in Figure 1.1. These properties have made the MOSFET technology the mainstream in design of high performance integrated circuits.

MOSFET transistors are generally used as switching devices in digital circuits with close to zero off current and very large turn on current. In static CMOS topology, the steady state current of a logic gate is very small [4]. In analog applications MOSFET devices are employed as active devices generally biased in strong inversion (SI) to be able to operate at high frequencies and at the same time keep the noise level very low. On the other hand, subthreshold (or WI) MOSFET devices are suitable for ULP applications where the device current density is very low [5].

Since most of the circuit topologies that are developed in this work are based on subthreshold MOS devices, in the rest of this Chapter a very brief review on the subthreshold MOS devices and their modeling techniques will be presented.

\(^1\)MOS device operates in weak inversion (WI) when the channel underneath the gate is weakly inverted by absorbing carriers. When the channel is completely inverted, the device will be in strong inversion (SI). The region in between is usually called medium inversion (MI) [1].
1.2 Device Modeling

A profound background on device modeling is essential to design high performance circuits. This Section provides necessary background for the design and analysis that will be carried out through the rest of this work. Figure 1.2 shows the structure of NMOS and PMOS transistors which are the main building blocks for implementing CMOS integrated circuits.

1.2.1 I-V Characteristics

The issue of MOSFET modeling in subthreshold regime has been extensively addressed in [1], [4], [6], and [7]. The EKV model \(^2\), first presented in [7], is based on an interpolation approach which can be used for all different regions of operation of an MOS device. In this model, all the voltage levels are referred to the local substrate voltage (not to the source voltage of a MOSFET device as it is usual in

\(^2\)Enz-Krummenacher-Vittoz (EKV).
BSIM model [1]). This property is especially interesting in this work where the bulk of transistor is used frequently as the second gate (or back gate [1]) of a device to provide more flexibility in the design.

Based on EKV model, the drain current of an NMOS transistor can be calculated by [7]:

$$I_{DS} = 2n\mu_c C_{ox} U_F^2 \frac{W}{L_e} \left( \ln^2(1 + e^{\left(\frac{V_P - V_D}{2kT/q}\right)}) - \ln^2(1 + e^{\left(\frac{V_D - V_P}{2kT/q}\right)}) \right)$$  \hspace{1em} (1.1)

where:

- \( n \) is the subthreshold slope factor which depends on process parameters as well as biasing condition, and is usually between 1 and 1.5,
- \( \mu_c \) in \([m^2/(V\cdot s)]\), is the effective carrier mobility in the channel and is different for electrons and holes:
  $$\mu_0 = A \cdot e^{B\sqrt{N_{ch}}}$$  \hspace{1em} (1.2)

where \( N_{ch} \) represents the channel doping density. For NMOS devices: \( A=1150 \), and \( B=-5.34\times10^{-10} \), and for PMOS devices: \( A=317 \), and \( B=-1.25\times10^{-9} \). Carrier mobility, also depends on Electric field.
- \( C_{ox} = \epsilon_{SiO_2}/t_{ox} \) is the gate oxide capacitance per unit area, \( \epsilon_{SiO_2} = K_{SiO_2}\epsilon_0 \) is dielectric constant of \( SiO_2 \), \( \epsilon_0=8.8541878176\times10^{-12} \) Fm\(^{-1} \), \( K_{SiO2}=3.9 \), and \( t_{ox} \) is the oxide thickness,
- \( U_T = kT/q \) is the thermodynamic voltage, \( k=1.3806504\times10^{-23} \) Jk\(^{-1} \) is Boltzmann’s constant\(^3 \), \( T \) is absolute junction temperature, and \( q=1.602\times10^{-19} \) C is the elementary electron charge,
- \( W \) and \( L_e \) are the effective channel width and length of the device,
- \( V_P \) is the device pinch off voltage.

The first part in (1.1) is called forward channel current, \( I_F \), and the second part is called reverse channel current, \( I_R \). Also, specific current of the device is defined as: \( I_S = 2n\mu_c C_{ox} U_F^2 \).

To complete the calculations using (1.1), it is necessary to calculate the values of \( V_P \) and \( n \). The pinch off voltage depends on the gate voltage \( V_G \) as:

$$V_P = V_G - V_{T0} - \gamma \cdot \left( \sqrt{V_G - V_{T0}} + \left(\frac{\Psi_0 + \frac{\gamma}{2}}{2}\right)^2 - \left(\frac{\Psi_0 + \frac{\gamma}{2}}{2}\right) \right)$$  \hspace{1em} (1.3)

where, \( V_{T0} \) stands for the device threshold voltage and is equal to the gate voltage when the inversion charge density in the channel \( (Q'_{inv}) \) is zero, or \( [4] \)

$$V_{T0} = V_{FB} + \Psi_0 + \gamma \sqrt{\Psi_0}$$  \hspace{1em} (1.4)

where \( V_{FB} \) is the flat band voltage, and

$$\gamma = \sqrt{\frac{2e_\epsilon N_{ch}}{C_{ox}}}$$  \hspace{1em} (1.5)

is the substrate factor or body effect, \( \epsilon_\epsilon = K_{Si}\epsilon_0 \) is the Si dielectric constant \( (K_{Si}=11.7) \), \( N_{ch} \) is the doping concentration in the substrate, \( \Psi_0 = 2\Phi_F + mU_T \) is the surface potential\(^5 \), \( \Phi_F = U_T \ln \left( N_{ch}/n_i \right) \) is the substrate Fermi potential, and \( n_i \) stands for the intrinsic carrier concentration of Si\(^6 \).

\(^3\)Although this coefficient is called by the name of Austrian physicist, Ludwig Boltzmann, it has been first introduced by German scientist Max Planck, in his derivation of the law of the black body radiation in December 1900 (see: \[9\]), and also \textit{http://www.wikipedia.org}.

\(^4[C]\) \( \equiv [A-s] \)

\(^5\)In this equation, \( m \) depends on the region of operation \([7]\).

\(^6\)\( \epsilon_{n_i} = 3.1 \times 10^{16} T^{3/2} \exp(-\frac{2002}{T}) \).
derivation of the gate voltage with respect to the pinch off voltage is defined as the device subthreshold slope factor given by:

\[ n \equiv \frac{dV_G}{dV_P} = 1 + \frac{\gamma}{2\sqrt{\psi_0 + V_P}} \]  

(1.6)

which can be simplified to:

\[ \frac{1}{n} = 1 - \frac{\gamma}{2\sqrt{V_G - V_{T0} + (\gamma/2 + \sqrt{\psi_0})^2}} \]  

(1.7)

It can also be shown that:

\[ V_P \approx \frac{V_G - V_{T0}}{n} \]  

(1.8)

In SI, (1.1) can be simplified to:

\[ I_{DS} \approx \frac{n \mu_e C_{ox}}{2} \frac{W}{L_e} \left( (V_P - V_S)^2 - (V_P - V_D)^2 \right) \]

\[ \approx \frac{\mu_e C_{ox}}{2n} \frac{W}{L_e} (V_G - nV_S - V_{T0})^2. \]  

(1.9)

It is noticeable that the current sensitivity to the source voltage is \( n \) times more than gate voltage. In other words: \( g_{ms} = n \times g_m \). Assuming that \( n \) is equal to one, then (1.9) can be simplified to the conventional equation.

In WI:

\[ I_{DS} \approx 2n \mu_e C_{ox} \frac{W}{L_e} U_T^2 \left( e^{\frac{V_P - V_S}{U_T}} - e^{\frac{V_P - V_D}{U_T}} \right). \]

\[ \approx 2n \mu_e C_{ox} \frac{W}{L_e} U_T^2 \left( e^{\frac{V_P - V_S}{U_T}} - e^{\frac{V_P - V_D}{U_T}} \right). \]  

(1.10)

where all the voltages are referred to the substrate and \( V_{T0} \) is independent to the \( V_{SB} \). In this work, we are frequently using (1.1), (1.9) and (1.10) for analysis purposes.

1.2.2 Second Order Effects

Mobility Reduction Due to Vertical Field

By increasing the vertical electric field, the carriers tend to flow closer to the silicon-oxide interface which results in more carrier scattering and hence mobility degradation. To include the effect of mobility degradation due to the vertical electric field, mobility can be replaced by the following value:

\[ \mu_e = \frac{\mu}{1 + \theta \cdot V_P} \]  

(1.11)

where \( \theta \) is a constant coefficient between 0.1 to 1 \( V^{-1} \) [7] [10]. A very approximate value for \( \theta \) is:

\[ \theta \approx 2 \times 10^{-9} / t_{ox} \]  

which shows more degradation for thinner gate oxides [10].

Velocity Saturation

Carrier mobility is proportional to the electric field by \( v = \mu E \) or more precisely [4]:

\[ v = \frac{\mu_e E}{\sqrt{1 + (\frac{E}{E_C})^n}} \]  

(1.12)

where \( n=1 \) for electrons and \( n=2 \) for holes, and \( E_C = v_{sat}/\mu_e \) is called critical electric field. In high electric field values when it becomes comparable to \( E_C \), the carrier velocity saturates due to the

---

7Experimental results in this work show that when one of the junctions in the MOS device becomes forward bias, this equation will not be precise enough. Using a modified substrate doping concentration can solve the problem. The other possibility is adding a bipolar device to the proposed MOS device in a proper configuration (see Chapter 2).
scattering phenomena. The scattering of carriers by high-energy phonons is the main reason for this speed limitation. In silicon, the carrier speed saturates at about \( v_{\text{sat}} = 10^5 \, \text{[m/s]} \) when the electric field approaches to about \( E_{\text{sat}} \approx 10^6 \, \text{[V/m]} \) [10].

As the device current depends on carrier velocity, this effect is generally modeled as the following [7]:

\[
I_{DS} = \frac{I_{DS0}}{1 + \frac{V_D - V_{Dsat}}{E_{sat}}}. \tag{1.13}
\]

Here, \( V_D^* \) is equal to \( V_D \) for triode MOS (\( V_D < V_{Dsat} \)), and equal to \( V_{Dsat} \) for saturated MOS (\( V_D > V_{Dsat} \)). Also, \( I_{DS0} \) is the current calculated without velocity saturation effect.

One of the main issues with the velocity saturation is that in short channel devices where \( V_{Dsat} \) is becoming larger than \( LE_{sat} \), the device current approaches to:

\[
I_{DS} \approx \mu C_{ox} \frac{W(V_G - nV_S - V_{T0})E_{sat}}{2}. \tag{1.14}
\]

which does not depend on channel length. In this case, the saturation voltage can be approximated by [4]

\[
V_{DSat} = \sqrt{\frac{2v_{sat}L_e(V_G - V_{T0})}{n\mu_e}}. \tag{1.15}
\]

As can be seen, based on (1.14), the quadratic relationship between current and voltage is modified to a first order linear equation. Generally, the relationship between current and voltage in strong inversion is modeled with an equation with the order of \( 1 < \alpha < 2 \).

Channel Length Modulation

When drain voltage is larger than the pinch off voltage, pinch off point starts to move towards the source and hence the channel length will be reduced by \( \Delta L \). Therefore, the drain current will be increased proportional to the channel length reduction as [4]:

\[
I_{DS} = \frac{I_{DS0}}{1 - \frac{\Delta L}{L}}. \tag{1.16}
\]

The channel length reduction can be calculated by [7]:

\[
\Delta L \approx \zeta \cdot \sqrt{V_D - V_P}. \tag{1.17}
\]

where

\[
\zeta = \frac{2\epsilon_S}{\sqrt{\gamma C_{ox}}} = \sqrt{\frac{2\epsilon_S}{qN_{ch}}}. \tag{1.18}
\]

Generally a simplified model for channel length modulation is used. In this approach, a resistance (output resistance) is put parallel to the drain-source of an MOS device. The value of this resistance can be calculated using \( g_{ds} \approx \frac{I_{DS}}{(\lambda \cdot L)} \). This approach is similar to introducing Early voltage in bipolar transistors where in MOS devices, Early voltage can be defined as: \( V_A = \lambda \cdot L \). By increasing the channel length or reducing the bias current, the parasitic effect of the channel length modulation can be reduced.

1.3 Design Considerations in Subthreshold

In this Section, some of the main issues associated with the MOS devices biased in subthreshold regime, such as variability, noise, and matching are addressed very briefly. As will be seen later, these nonideality effects can increase the design cost in terms of area, energy consumption, and reliability.
CHAPTER 1. SUBTHRESHOLD MOS FOR ULTRA-LOW POWER

1.3.1 PVT Variation

Rewriting (1.10) in the form of:

\[ I_{DS} \approx I_0 e^{\frac{V_G}{nU_T}} (e^{-\frac{V_S}{nU_T}} - e^{-\frac{V_D}{nU_T}}) \]  

(1.19)

it clearly illustrates the exponential I-V characteristics of a MOS device biased in WI (subthreshold) regime. This characteristic is on one hand useful for implementing widely tunable circuits, while on the other hand, it represents the high sensitivity of the circuit to PVT variations. For example, any small variation on device threshold voltage \( V_{T0} \) will be translated to exponential variation on the bias current.

It is also instructive to calculate the temperature dependence of the bias current in subthreshold regime. Assuming \( \mu = \mu_0 \left( \frac{T}{T_0} \right)^\alpha \):

\[ \frac{\partial I_{DS}}{\partial T} \approx I_{DS} \cdot \left( \frac{\alpha + 2}{T} - \frac{\partial V_{T0}/\partial T - V_{T0}/T}{nU_T} \right). \]  

(1.20)

To derive this equation, the temperature dependence of subthreshold slope factor has been ignored. Meanwhile, it is assumed that \( V_S << U_T \) and \( V_D >> U_T \) which is not the case for all the possible configurations. Based on (1.20), it is possible to show that in WI:

\[ I_{DS} = I_{DS0} \cdot e^{\frac{F}{\eta}} \cdot \left( \frac{T}{T_0} \right)^G \cdot e^{-\frac{F}{T}} \]  

(1.21)

where, \( G = \alpha + 2 - \theta q/(nk) \), \( F = qV_{T0}/(nk) \) which is independent of temperature, and \( V_T \approx V_{T0} + \theta(T - T_0) \). On the other hand in SI, the temperature variation of the device current can be calculated by:

\[ I_{DS} = I_{DS0} \cdot \left( \frac{T}{T_0} \right)^\alpha \cdot \left( \frac{V_G - nV_S - V_{T0} - \theta(T - T_0)}{V_G - nV_S - V_{T0}} \right)^2. \]  

(1.22)

The thermal variation of the bias current is depicted in Figure 1.3. As illustrated in this figure and can be concluded from (1.21) and (1.22), by moving toward subthreshold region, the variations due to the temperature increases rapidly.

---

9 Here, \( T_0 \) is the temperature in which \( \mu_0 \) has been measured. Meanwhile, \( \alpha \) is equal to -2.4 for electron and -2.2 for hole in Si [1].

10 Here, it is assumed that the threshold voltage linearly depends on temperature and the proportionality factor is \( \theta \) [4].
1.3.2 Matching

Device mismatch is one of the most important design issues especially in design of high performance analog and digital systems in modern ultra-deep-submicron (UDSM) technologies. Experiments show that the two main sources of introducing mismatch among devices are difference in threshold voltage ($\Delta V_T$) and current factor ($\Delta \beta$, where $\beta = \mu C_{ox} W/L_e$). The difference among devices raised from difference in $V_T$ and $\beta$ have random nature with a normal distribution where their mean values are $V_{T0}$ and $\beta_0$ [11]. The variance of these parameters can be presented by

$$\sigma^2(\Delta V_T) = \frac{A_{VT}^2}{W \cdot L}$$

(1.23)

$$\left(\frac{\sigma(\Delta \beta)}{\beta}\right)^2 = \frac{A_{\beta}^2}{W \cdot L}$$

(1.24)

where proportionality constants $A_{VT}$ and $A_{\beta}$ are technology dependent parameters.

For simple current mirrors and differential pair configurations, it can be shown that the mismatch between current values and input referred voltage offset are, respectively:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 = \left(\frac{\sigma(\Delta \beta)}{\beta}\right)^2 + \left(\frac{g_m}{I}\right)^2 \sigma^2(\Delta V_T)$$

(1.25)

$$\sigma^2(\Delta V_{GS}) = \sigma^2(\Delta V_T) + \left(\frac{I}{g_m}\right)^2 \left(\frac{\sigma(\Delta \beta)}{\beta}\right)^2$$

(1.26)

Since the value of $g_m/I$ has its maximum value in WI [Figure 1.5], and regarding (1.25) and (1.26), it is expected that the voltage matching improves slightly by moving towards WI,\(^{11}\) while the current matching degrades. This implies that implementing current mirrors with acceptable level of matching will be much more difficult in WI region compared to the current mirrors implemented in SI region.

![Figure 1.4: Expected offset voltage at the input of a differential pair circuit by technology scaling when minimum size devices are utilized. Data values are extracted from [11].](image)

Figure 1.4 shows the expected value of the input referred offset of an NMOS differential pair circuit by technology scaling. Although the value of $A_{VT}$ and $A_{\beta}$ are improving by technology scaling, however, the size of devices are reducing as well, and consequently the expected offset value increases considerably. Depicted in Figure 1.4, the input referred offset increases by a factor of about 12mV/decade by technology scaling.

Physical Mechanism of $V_T$ Fluctuation

Threshold voltage of an MOSFET device can be expressed by:

$$V_T = V_{FB} + 2\phi_B + \frac{Q_d}{C_{ox}}$$

(1.27)

\(^{11}\)Generally the term which depends on $V_T$ variation is dominant over the term depending on the variations due to $\beta$. Therefore, the expected reduction on input referred offset voltage is not considerable.
where $Q_d$ is the depletion layer charge and $\phi_B$ is the surface potential. Based on this, any variation on channel doping concentration, surface state charge density ($Q_{ss}$), and gate oxide thickness can result in variation on the device threshold voltage. The variation on surface potential, $\delta \phi_B$, can be estimated by $\delta \phi_B \approx U_T \cdot \delta N_A / N_A$, where $\delta N_A$ is the fluctuation on substrate doping [12]. It can be shown that threshold voltage fluctuation due to the random dopant fluctuation (RDF) can be estimated by [12]:

$$\sigma_{V_T} = \frac{\sqrt{q^3 \epsilon Si \phi_B}}{\sqrt{2 \epsilon_{ox}}} \cdot t_{ox} \cdot \frac{1}{\sqrt{W_{eff}(L_{eff} - W_d)}}$$  \hspace{1cm} (1.28)$$

where $W_d$ is the average of the maximum pn junction depletion layer width of the drain $n^+$ region. This expression indicates that threshold voltage fluctuation increases approximately by a factor of $\sqrt{\kappa}$ by technology scaling where $\kappa > 1$ is the scaling factor under constant field scaling rule [12]. Some more recently published reports are proposing this expression for standard deviation of device threshold voltage [13]:

$$\sigma_{V_T} = 3.19 \times 10^{-8} \cdot t_{ox} \cdot 2 \sqrt{N_A} \cdot \frac{1}{\sqrt{W_{eff}L_{eff}}},$$  \hspace{1cm} (1.29)$$

which indicates a weaker dependence on channel doping concentration compared to (1.28).

To prevent the increase on threshold voltage variation with technology scaling, $V_T$ adjustment method needs to be modified. For example, instead of controlling the depletion layer charge, new gate materials could be used to avoid increasing substrate doping concentration. There are other sources for increasing the threshold voltage variability, such as line edge roughness and oxide thickness variation. While the effect of line edge roughness can be neglected, the variation of threshold voltage due to the oxide thickness is about half of the variation due to the RDF [14].
Mismatch due to Gate Leakage

The gate leakage current adds a new source of device mismatch which should be included in calculations. The variation on drain current including the gate leakage mismatch is [15]:

\[ \sigma_{I_{DS}}^2 \approx \left( \frac{AVT}{\sqrt{WL}} \cdot \frac{g_m}{I_{DS}} \right)^2 + \left( \frac{X_{IGS} \cdot I_G}{\sqrt{WL}} \right)^2 \]

where \( X_{IGS} \approx 0.03 \).

1.3.3 Noise

The model that is used generally to model the MOS device noise includes drain thermal noise and gate voltage flicker noise is

\[ i_{n,d}^2 = 4kT \gamma g_m \]

\[ v_{n,f}^2 = \frac{4kT \rho}{WL} \cdot \frac{1}{f^\alpha} = \frac{k_f}{WL \cdot 1f} \cdot \frac{1}{f^\alpha} \]

Flicker noise is inversely proportional to the frequency \( f \), and \( k_f = 4kTC_{ox} \rho \) [6]. The empirical coefficient \( k_f \) for NMOS devices is essentially independent of bias, fabricator and technology (\( k_{f,NMOS} = 10^{-24} \)), while for PMOS devices, this coefficient is smaller\(^{12}\) and depends on biasing condition [16]. To reduce the effect of flicker noise, the most effective way is to increase the device dimensions [17].

To have a unified thermal noise model for SI and WI regions, \( \gamma \) (excess noise factor) has been defined as the following in [6]:

\[ \gamma = \frac{2n}{3 + \left( \frac{2nU_T}{I_{DS}} \right)^2} \]

which results in \( \gamma = n/2 \) in WI and \( \gamma = 2n/3 \) in SI. The thermal noise power-spectral density can also be interpolated from WI to SI using the following function:

\[ \frac{1}{R_N} = g_m \times \frac{1}{1 + i_f} \cdot \frac{1 + \alpha}{2} \cdot \frac{2}{3} \cdot i_f \cdot \frac{1 + \sqrt{\alpha} + \alpha}{1 + \sqrt{\alpha}} \]

where \( \alpha = \frac{i_r}{i_f} \), \( i_f \) and \( i_r \) are the forward and reverse currents in the channel normalized to specific current \( I_S \equiv 2nU_T^2 \),\(^{13}\) and \( R_N \) is the equivalent noise resistance of the channel (\( v_n^2 = 4kTR_N \)). It is interesting to notice that the channel noise increases when device moves from saturation (\( \alpha=0 \)) to conduction (\( \alpha=1 \)). Also, the channel noise increases slightly when device moves from WI (low \( i_f \) values) toward SI (high \( i_f \) values).

In [1], the channel thermal noise has been calculated as

\[ i_{n,d}^2 = 2qI^* \left( 1 + e^{-\frac{V_{DS}}{I^*}} \right) \]

where \( I^* \) is the current in flat part of the \( I_{DS}-V_{DS} \) curve (or in other words: \( V_{DS} > 5U_T \)). Although, this expression has been derived assuming the presence of thermal noise in channel, it is corresponding to shot noise associated with the dc flow produced by carriers crossing the source-channel barrier [1]. It is also noticeable that the current noise increases with reducing the \( V_{DS} \).

In very high frequencies, where the transient time of carriers between source and drain becomes important, a new source of noise should be added to the MOS device model. The finite carrier transition time in the channel adds a positive term or equivalently a resistive part to the input impedance of an\(^{12}\)\( k_{f,PMOS} \) can be 50 times smaller than \( k_{f,PMOS} \)\(^{17}\).

\(^{13}\)Forward and reverse currents can be calculated from (1.1) where the first term stands for forward current and the second term stands for reverse current.
MOS device. The noise associated with this effect can be modeled by a noise current source at the gate with mean-square power of [10]:

\[ \dot{i}_{n,g}^2 = 4kT\delta g_d \Delta f = 4kT\delta \cdot \frac{\omega^2 C_{gs}^2}{5g_{ao}} \cdot \Delta f \]  

(1.36)

where \( \delta \) is typically 4/3. This noise is correlated with the drain thermal noise with correlation factor of 

\[ c \equiv \frac{i_{n,g} \cdot i_{n,d}^*}{\sqrt{i_{n,g}^2 \cdot i_{n,d}^2}} = j0.395 \]  

Noise Efficiency Factor

To be able to compare the noise performance of a specific design with other designs, noise efficiency factor (NEF) has been defined in [17]. For this purpose, the total equivalent input noise of an ideal bipolar transistor (including only thermal noise without considering the base resistance noise) has been defined to be the reference noise level:

\[ v_{rms,in,bip} = \sqrt{BW \cdot \frac{\pi}{2} \cdot \frac{4kT}{g_m}} \]  

(1.37)

where \( g_m = I_C/U_T \) in a bipolar transistor (\( I_C \) is the collector current). Also, \( BW \) represents the circuit bandwidth. In case of a simple bipolar transistor, the bandwidth is \( f_t \) (transient frequency of a bipolar transistor or the frequency at which the current gain of transistor becomes one) [17]. To calculate the NEF for a circuit with equivalent input referred noise of \( v_{rms,in} \):

\[ NEF = \frac{v_{rms,in}}{v_{rms,in,bip}} \]  

(1.38)

For example, for a simple MOS transistor in SI:

\[ v_{rms,in,MOS}^2 = BW \cdot \frac{\pi}{2} \cdot \frac{4kT}{g_m} = BW \cdot \frac{3kT(V_{GS} - V_T)}{I_{DS}} \]  

(1.39)

Assuming that the device is operating on the boundary of SI (i.e., \( V_{GS} - V_T = 2\sqrt{10}nU_T \)), then \( NEF = 2.43 \) [17]. Therefore, equivalent noise of a CMOS design in SI with the same amount of power dissipation and bandwidth is about five times more than a bipolar design. In [18]-[22] some techniques for implementing amplifiers with very low NEF values have been reported. Chopper stabilization have been used in [20] to reduce the flicker noise and the offset voltage. In [21], careful current partitioning technique has been used to improve the NEF to 3.81 in a folded-cascode operational transconductance amplifier (OTA). To reduce the NEF to 1.8, partial OTA sharing technique has been introduced in [22]. In this design, large size devices have been used to make the flicker noise effect negligible.

Noise due to the Gate Leakage

The noise of gate leakage current is a shot noise as the noise in other types of PN junctions. The noise current density can be expressed by [15]

\[ S_{IG} = 2qI_{GS} \]  

(1.40)

that should be included in the estimation of circuit noise. To calculate the gate leakage current [15]:

\[ I_{GS} = A \cdot V_{INV} \cdot V_{GS} \cdot e^{B \cdot V_{GS}} \]  

(1.41)

which represents exponential dependance of the gate current on the gate-source voltage of a device. In this equation:

\[ A = \frac{I_{GINV} \cdot e^{-\frac{B}{2} \cdot V_{INV}}}{2} \]  

(1.42)
and
\[ B = \frac{3}{8} \cdot \frac{B_{INV}}{X^2_B} \] (1.43)
and
\[ V_{INV} = nU_T \cdot \ln \left( 1 + e^{\frac{V_{GS}-V_T}{nU_T}} \right) \] (1.44)

Here, \( X_B \) is the oxide potential barrier which is 3.1V for electrons and 4.5V for holes. \( I_{GINV} \) and \( B_{INV} \) are physical parameters depending on \( t_{ox} \), \( L \), and \( W \). For electrons:

\[ I_{GINV} = 1.6 \cdot 10^{-4} \cdot \frac{WL}{t_{ox}^2} \] (1.45)
and
\[ B_{INV} = 2.9 \cdot 10^{10} \cdot t_{ox} \] (1.46)

These values can be replaced in (1.41) to estimate the gate leakage current.

### 1.4 Ultra-Low-Power Design Using Subthreshold MOS

Using subthreshold MOS devices for implementing low-voltage and very low-power analog and digital circuits can be traced back to 70’s \[5\], \[23\]. While in most of the applications at that time, MOS devices were employed in strong-inversion, the need for reducing the power consumption and supply voltage encouraged the designers to develop special design techniques for using subthreshold MOS devices. Some industrial applications such as low-power quartz wristwatches \[24\] promoted even more the researchers to establish the required bases to simplify and increase the reliability of using subthreshold MOS devices. For this purpose, many different design and device modeling techniques have been proposed \[5\], \[7\].

In \[23\] at 1970, it was shown that it is possible to reduce the supply voltage of a CMOS inverter down to \( V_{DD} \approx 4U_T \) with sufficient gain for logic operation. Therefore, it is possible to use CMOS logic circuits deeply biased in subthreshold regime. This means that when the speed of operation is not the premier design issue, it is possible to reduce the supply voltage and hence reduce the power dissipation of the system which is mostly proportional to the dynamic power consumption.

Afterwards, the concept of low-power design using reduced supply voltage has been developed even more to construct more complex integrated circuits with possibility of dynamic power management \[25\]. In this type of systems, supply voltage can be scaled in a very wide range to minimize the power dissipation with respect to the operation frequency or work load \[26\].

Figure 1.6 shows the trends in semiconductor industry based on the data points and predictions made in 2001 \[27\]. All the parameters in these two graphs are normalized to their nominal values in the year 2001. While device channel length (\( L \)) has been scaled down progressively, the scaling for supply voltage, \( V_{DD} \), and gate oxide thickness has not been as aggressive as scaling of channel length. As illustrated in these graphs, there is a very rapid increase in the static power consumption that becomes more and more pronounced in more advanced technology nodes. Therefore, to design ULP systems in modern technologies, special cares are required to overcome this problem.

Emerging new applications that require very low power consumption, has made subthreshold circuits very popular. In these type of applications, energy consumption and cost are the most important parameters with medium (1Msps - 10Mps) or low (10ksps - 100ksps) data throughput systems \[14\]. Lowering the supply voltage even below threshold voltage of devices leads to quadratic reduction of the circuit dynamic power. This technique also is helpful to reduce the leakage or static power consumption of conventional CMOS circuit topologies implemented in modern nano-scale technologies.
In the following, the two main issues in design of ultra-low power digital circuits, i.e. static power dissipation and variability will be reviewed. In more advanced deep-submicron MOS technologies, these two problems are more pronounced. Therefore, if not necessary, generally older technologies can be used for implementing energy-constrained circuits that does not require a high performance, such as in RFIDs, bio-implants, and sensor network. In some applications, the energy-constrained circuit needs to have a high performance while occasionally is operational [28]. In this case, dynamic voltage scaling can be employed to scale the circuit power consumption and performance by moving from subthreshold region to superthreshold (above threshold) region. In such bursty applications, an advanced CMOS technology needs to be used to support the required specifications during high performance mode of the operation [28]. Advanced MOS technologies also have been used for implementing energy-constrained circuits which are supporting a high-performance application. In these cases, special design techniques are required to implement subthreshold circuits which suffer from high leakage current and very wide parameter variability [28].

1.4.1 MOS Transistor Leakage Mechanisms

While the static power consumption of static CMOS circuits have been ignored in early CMOS technologies [29], it has become a major challenge in UDSM technologies. Figure 1.7 describes the main leakage mechanisms in a deep sub-micron MOS device. Among different types of leakage, subthreshold leakage current and gate tunneling currents are more essential. The main sources of static power consumption in CMOS logic circuits that are more pronounced in modern technologies are briefly explained in this Section (see also: [30], [31], [32]).

Scaling Rules

To keep the transistor performance on an acceptable level, in addition to scaling the device length, \( L \), it is necessary also to scale gate oxide thickness, \( t_{ox} \), junction depth, \( X_j \), and depletion depth, \( D \). This proportional scaling results in an acceptable device aspect ratio defined by

\[
AR = \frac{L}{\sqrt{t_{ox}X_jD_{eox}}}. \tag{1.47}
\]

Unfortunately, it is difficult to keep the device \( AR \) on an acceptable level in very deep sub-micron technologies. Specially, maintaining the vertical sizes on desired value is very difficult. As will be seen
in the next Section, when gate oxide approaches scaling limits, there is a rapid increase in gate oxide leakage. Therefore, it is difficult to scale down the gate oxide thickness as device channel length. This will prevent having appropriate device \( AR \).

**Gate Tunneling**

Oxide leakage is due to tunneling through the gate oxide. In more advanced technologies where oxide thickness, \( t_{ox} \), is reducing and hence the field across the oxide is more, the tunneling phenomena becomes more significant. The gate tunneling current is due to the two different mechanisms: Follwer-Nordheim (FN) tunneling, and direct tunneling. The FN tunneling current density is given by \[4\]

\[
J_{FN} = \frac{q^3 E_{ox}^2}{16\pi^2\hbar\phi_{ox}} \exp \left( -\frac{4\sqrt{2m^*\phi_{ox}^3}}{3hqE_{ox}} \right)
\]

(1.48)

where \( E_{ox} \) is the field across the oxide, \( \phi_{ox} \) is the effective height for electron in the conduction band, and \( m^* \) is the effective mass of an electron in the conduction band of silicon. On the other hand, the current density of the direct tunneling is \[4\], \[33\]

\[
J_{DT} = \frac{q^3 E_{ox}^2}{16\pi^2\hbar\phi_{ox}} \exp \left( -\frac{4\sqrt{2m^*\phi^3}}{3hqE_{ox}} \cdot (1 - \sqrt{1 - \frac{V_{ox}}{\phi_{ox}}})^2 \right)
\]

(1.49)

By reducing the gate oxide thickness, the direct tunneling current increases rapidly.

In analog applications, it is possible to model the gate leakage current by a conductance \( g_{tun} \) in parallel to the gate capacitance \( C_g \). In frequencies higher than \( f_g = g_{tun}/(2\pi C_g) \), the input impedance is capacitive while for frequencies lower than \( f_g \), it is resistive. As shown in \[15\], the gate cutoff frequency can be calculated by

\[
f_g = \frac{g_{tun}}{2\pi C_g} \approx A \cdot V_{GS}^2 \cdot e^{t_{ox}(V_{GS}-13.6)}
\]

(1.50)

where \( t_{ox} \) is in \([\text{nm}]\) unit, and \( A \) is a constant number (1.5\(\times\)10\(^{16}\) for NMOS transistors and 0.5\(\times\)10\(^{16}\) for PMOS devices). When \( f_g \) is about 0.1Hz for 0.18\(\mu\)m CMOS, it increases to about 1MHz in 65nm CMOS \[11\].

---

*Figure 1.7: Leakage current sources in a MOS device.*

**Draft version of the book. October 26, 2009**
Subthreshold Conducting

Subthreshold (weak inversion) conducting current is due to the drift of minority carriers at \( V_{GS} < V_{TH} \). The minority carrier concentration in this region of operation is very low but not zero. The weak inversion current can be estimated using (1.10) where:

\[
\text{n} = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{\epsilon_{Si}}{\epsilon_{ox}} \cdot \frac{t_{ox}}{W_{dm}}. \tag{1.51}
\]

Here, \( W_{dm} \) is the maximum depletion region width, and \( C_{dm} \) is the capacitance of the depletion region [4]. The leakage current due to the subthreshold current is generally characterized by the subthreshold slope:

\[
S = \left( \frac{d \log_{10} I_{DS}}{dV_{GS}} \right) = 2.3nU_T = 2.3U_T \left( 1 + \frac{t_{ox}}{W_{dm}} \cdot \frac{\epsilon_{Si}}{\epsilon_{ox}} \right). \tag{1.52}
\]

Subthreshold slope indeed represents how effectively the transistor can be turned off when \( V_{GS} \) is decreased below threshold voltage. As it is illustrated in Figure 1.8, a lower subthreshold slope results in smaller off current, \( I_{OFF} \). Higher value for \( V_T \), helps to reduce the off current. However, high \( V_T \) devices (HVT) results in lower on current, \( I_{ON} \), and hence increased gate delay.

![Figure 1.8: I-V characteristics of an NMOS transistor and effect of subthreshold slope factor on off current of the device.](image)

PN junction

Reverse biased pn junction leakage has two main components: minority carrier diffusion and drift near the edge of the depletion region, and the other one is due to the electron-hole pair generation inside the depletion region of reverse-biased pn junction. When the p-side and n-side of the junction are heavily doped, which is the case in MOSFET devices, then band-to-band tunneling current should be added to the estimations. The tunneling current density is given by [4]

\[
J_{b-b} = \frac{AEV_R}{\sqrt{E_g}} \exp \left( -B \sqrt{\frac{E^3}{E}} \right) \tag{1.53}
\]

where \( A = \sqrt{2m^*q^3/(4\pi^3h^2)} \), and \( B = 4\sqrt{2m^*/(3qh)} \), \( m^* \) is effective mass of electron, \( E_g \) is the energy bandgap, \( V_R \) is the applied reverse biased voltage, \( E \) is the electric field at the junction, and \( h = \hbar/(2\pi) \), and \( \hbar = 6.62606896 \times 10^{-34} \) [J.s] is Planck’s constant. Assuming a step junction, the electric field can be calculated by

\[
E = \sqrt{\frac{2qN_aN_d(V_R + V_b)}{\epsilon_{Si}(N_a + N_d)}} \tag{1.54}
\]

where \( N_a \) and \( N_d \) are the doping concentration in the p and n side of the junction.
Body Effect

Sometimes designers apply a voltage to the body terminal of a device to control the leakage current or speed of operation. This body bias voltage can be positive (to reduce the threshold voltage and hence speedup the circuit), or negative (to increase the threshold voltage and reduce the leakage current). To reduce the leakage current, one possible approach is to increase the threshold voltage by reverse biasing bulk-to-source junction and widening the bulk depletion region:

$$V_T = V_{FB} + 2\psi_B + \sqrt{2\varepsilon_S q N_{ch} (2\psi_B + V_{BS})}$$  \hspace{1cm} (1.55)

where $\psi_B = U_T \ln \left( N_{ch}/n_i \right)$. However, this will increase the sensitivity of the threshold voltage to the bulk voltage, which can be estimated by [31]

$$\frac{\partial V_T}{\partial V_{BS}} = \frac{1}{C_{ox}} \sqrt{\frac{\varepsilon_S q N_A}{2(2\psi_B + V_{BS})}}.$$  \hspace{1cm} (1.56)

The sensitivity will be higher for higher reverse voltages.

DIBL

Drain voltage can affect the channel charge like gate voltage, especially in very short-channel devices. In short-channel devices because of proximity of the source and drain, drain voltage can influence the depletion region beneath the channel and hence change the channel potential. Drain-induced barrier lowering (DIBL) affects the leakage current by reducing the effective device threshold voltage [4]. In short-channel devices, the source-drain potential have a considerable effect on band bending over the channel. Therefore, the threshold voltage and consequently the subthreshold current of device can vary with this voltage. Indeed, in short-channel devices the depletion region of source and drain junctions interact to each other near the channel surface and will reduce the potential barrier between source and drain. Higher drain voltage or shorter channel length with enhance the DIBL effect. DIBL generally happens before the pinchthrough via the bulk occurs [31].

DIBL does not change the subthreshold slope. To reduce the effect of DIBL, higher surface and channel doping and shallow source and drain junction depths are required. The DIBL coefficient, $\eta$, can be expressed as [34]

$$\eta = \frac{1}{2 \cosh \frac{L_{eff}}{2L_t}}$$  \hspace{1cm} (1.57)

in which $L_t$ is a characteristics length:

$$L_t = \sqrt{\frac{\varepsilon_S t_{ox} W_{dm}}{\epsilon_{ox} K}}$$  \hspace{1cm} (1.58)

and $K$ is a fitting parameter. Based on this expression, by scaling the transistor length, DIBL coefficient is increasing.

The bias current of an MOS device biased in subthreshold regime including DIBL and body effect can be modeled by [31]

$$I_{DS} = I_{DS0} \times e^{\frac{V_{GS} - V_{TF} - \Delta V_T}{\sqrt{2}} \times \left( 1 - e^{-\frac{V_{DS}}{V_T}} \right)}$$  \hspace{1cm} (1.59)

where:

$$I_{DS0} = \mu_0 C_{ox} \frac{W}{L_e} U_T^2 e^{-\frac{\Delta V_T}{V_T}}.$$  \hspace{1cm} (1.60)

Here, $\Delta V_T$ is added to consider the threshold voltage variation from one transistor to the other one. The exponential dependence of $I_{DS0}$ on $\Delta V_T$ shows the high sensitivity of the subthreshold current on process variation.
Regarding (1.59), the subthreshold leakage current could be calculated by

\[ I_{\text{sub}} \approx I_{DS0} \times e^{-\frac{V_{PD} + \gamma V_D}{N_{max}}} \times \left(1 - e^{-\frac{V_{PD}}{N_{max}}}ight) \] (1.61)

which is very sensitive to DIBL effect.

**GIDL**

Gate-induced drain leakage (GIDL) is due to the high electric field near the Si-SiO\(_2\) interface. The high gate-drain electric field can give sufficient energy to the electrons or holes to cross the interface potential barrier and enter into the oxide region and hence create a current flow between drain and substrate. To reduce the GIDL effect, very high and abrupt drain doping concentration which provides lower series resistance should be used [35].

**Hot Carrier**

Hot carrier injection is due to the high electric field near Si-SiO\(_2\) interface [4]. High electric field can give sufficient energy to the carriers to cross the interface potential barrier and enter into the oxide layer.

**Punchthrough**

Channel punchthrough is due to the proximity of the drain and source in short-channel devices [35]. In this case, the depletion region at the drain-substrate and source-substrate junction extend into the channel. This phenomena will reduce the effective channel length. Therefore, increasing the reverse bias voltage across the junctions by increasing V\(_{DS}\) pushes the junction closer to each other. Punchthrough happens when the depletion regions merge together [35].

**Channel Length Effect**

The threshold voltage reduction of an MOS device when the device length is reducing called threshold voltage rolloff [4]. The reduction of threshold voltage can be worsen in higher drain-source voltages. A nonuniform HALO doping can be used to mitigate this problem by reducing the depletion width and hence reducing the DIBL effect [36]. As a result, reverse SCE (RSCE) occurs and threshold voltage decreases by increasing the length of device [37].

**Narrow-Width Effect**

The threshold voltage of an MOS device also depends on the width of transistor [4], [31],[38], [39]. Depending isolation technologies, threshold voltage can be reduced or increased by reducing the channel width. With a less abrupt transition between the channel and the isolation, such as in local oxidation of silicon (LOCOS), the device threshold voltage increases with reducing the channel width. This effect is mainly because of extra depletion charge beneath the field oxide that should be added to the channel charge [32]. This effect is inverse for abrupt isolations such as in sealed interface local oxidation (SILO), and shallow trench isolation (STI).

**Thermal Effect**

The stand-by current of a transistor can change considerably by temperature. This variation is mainly due to carrier mobility (\(\mu\)), thermal voltage (U\(_T\)), subthreshold slope factor (\(n\)), and threshold voltage [32]. Subthreshold slope (\(S\)) increases with temperature almost linearly, while threshold voltage decreases with temperature (the coefficient is about -0.8mV/\(\circ\)C) [4].
Short Circuit Current

Because of finite transition time at the input of a static CMOS gate, during a very short period of time both PMOS and NMOS devices are on and hence there is a short circuit current between $V_{DD}$ and ground. This current can be considerable when $V_{DD}$ is high and both PMOS and NMOS devices conducting in SI. When the logic circuits are biased in subthreshold regime, this current can be ignored most of the time [32].

1.4.2 Leakage Reduction Techniques

The total power consumption of a digital system is the sum of dynamic ($P_D$), and leakage (or static) power consumption ($P_{leak}$) can be approximated by [30]

$$P_{diss} \approx P_D + P_{leak} \quad (1.62)$$

where

$$P_D = \alpha f_{op} C V_{DD}^2 \quad (1.63)$$

and

$$P_{leak} = I_{leak} \cdot V_{DD} \quad (1.64)$$

where $\alpha$ stands for the average switching activity rate. To control the static power consumption of CMOS logic circuits which is going to be more and more pronounced in advance technologies, special techniques are needed to be used [4], [31]. Some of these techniques are briefly explained in the following.

Device Level Engineering

The leakage current, as explained before, depends on different physical phenomena and can be reduced by controlling the device dimensions (such as length, $L$, oxide thickness, $t_{ox}$, junction depth, $X_j$), and doping profile of the transistor.

In device engineering level, it is very important to control the short-channel effects (SCEs) by scaling down the device dimensions and choosing proper channel doping profile. Generally it is very desirable to scale the device dimensions under constant field principle [4]. Using retrograde doping and halo doping are two possible approaches to control the SCEs [4].

Circuit Techniques

At the circuit level, it is possible to reduce the leakage current contribution through careful selecting voltage levels in different terminals of devices, and choosing proper devices with appropriate threshold voltages. Careful device sizing is the other possibility to reduce the leakage current. In many ultra-low power designs, the length of MOS devices is generally selected slightly larger than minimum size to reduce the leakage current and have less variability [14], [40], and [41]. It is also possible to use special circuit topologies to control the static current [32].

A common circuit technique that can be used for reducing the leakage current, as an example, is using stacked transistors (stacking effect). This technique, depicted in Figure 1.9, can reduce the leakage current by one order of magnitude compared to a single transistor configuration [42], [43]. The main issue associated with this technique is the dependence of leakage current on input data vector [44].

Multiple threshold voltage CMOS technologies (MTCMOS) provide this possibility to use different types of devices for different purposes. In other words, one can use HVT devices for reducing the leakage current and use LVT devices in critical paths where the speed of operation is important. To fabricate multiple threshold devices in a technology, it is possible to change the channel doping, oxide
CHAPTER 1. SUBTHRESHOLD MOS FOR ULTRA-LOW POWER

1.4.3 Variability

Variability is one of the major concerns in design of any type of integrated circuits. For a long time, designers have been trying to somehow reduce, control, or compensate the effect of process variation on circuit performance.

In the field of digital design, gate delay variation due to process variation has been always an important concern. This effect is more pronounced in subthreshold logic circuits where current of MOS devices exponentially depends on gate voltage, and threshold voltage. Therefore, any small variation on the device parameters can change considerably the peak current (device on current), and off current of the device, and hence change the gate delay, and also the static and the dynamic current consumption of the circuit [14], [47]. Figure 1.10 shows the effect of process variation on different device parameters in CMOS 65nm technology. As can be seen, by moving towards subthreshold regime (lower $V_{DD}$ values), the amount of variation on the cell turn on current, $I_{ON}$, and delay, $t_d$, increases rapidly. The variation on turn off current, $I_{OFF}$, is always high because this current is always determined by the subthreshold current. It is also noticeable that the ratio of turn on to turn off current, $\gamma$, degrades considerably by reducing the supply voltage.

In addition, process variation and device mismatch can degrade the circuit reliability. For example, device parameter variation can degrade the static noise margin (SNM) of memory or logic cells considerably [14], [28]. To compensate the effect of process or environmental variations, many different techniques have been proposed. A common approach for mitigating this effect is to use up-sized channel length devices which helps to reduce the variability and improve subthreshold factor, simultaneously [14]. The other possibility is to increase the circuit supply voltage to a high enough value to make sure that the circuit will remain operational even in presence of variation [28].
Variation Effect on Static Noise Margin

Static noise margin (SNM) is a measure for reliable operation of a logic cell [48], [49]. To explore the effect of process variation on logic cell operation, in this Section the SNM of an inverter will be analyzed. Since ULP applications are the main concern of this work, we are assuming that all the devices are biased in subthreshold regime. In other words, the circuit supply voltage is not more than threshold voltage of the MOS devices. Figure 1.11 shows a CMOS inverter and the corresponding Butterfly curve that can be used for measuring its SNM.

Figure 1.10: Variation on: (a) $I_{ON}$ current, (b) $I_{OFF}$ current, and (c) delay of a NAND gate implemented in 65nm CMOS technology. (d) $\gamma = I_{ON}/I_{OFF}$.

Figure 1.11: A sample CMOS inverter and the corresponding Butterfly curve used for estimating SNM.
With this assumption, the bias current of MOS devices could be estimated by:

$$I_{DS} = I_0 \cdot e^{\frac{V_{G} + V_{DD}}{nUT}} \cdot (1 - e^{-\frac{-V_D}{n\mu C_{ox} W L e^{V_S}}} \cdot (1 + \frac{V_D - V_S}{V_A}) \right)$$

(1.65)

where:

$$I_0 = 2n\mu e C_{ox} W L e^{V_S}$$

(1.66)

and $V_A$ represents the effect of finite output resistance. All voltages are refereed to the bulk of device. Although not necessary, in the rest of this Section it is assumed that the subthreshold slope factor, $n$, and $V_A$ values are equal for NMOS and PMOS devices in order to simplify the equations.

To maximize the typical SNM of the gate, generally the size of PMOS device is selected larger than NMOS transistor such that:

$$I_{0,NMOS} = I_{0,PMOS}.$$  

(1.67)

With this constraint, the crossover voltage will be as close as possible to $V_{DD}/2$ and logic cell will have a relatively symmetric rise and fall transitions.

Now, to calculate the voltage transfer characteristic (VTC) of the inverter, the following equation should be solved:

$$I_{DS,NMOS} = I_{SD,PMOS}$$

(1.68)

which results in:

$$K_\Delta \cdot e^{\frac{V_{DD}(1+\eta)}{n\mu C_{ox} W L e^{V_S}}} \cdot e^{\frac{2V_{DD}}{n\mu C_{ox} W L e^{V_S}}} \cdot \frac{1 - e^{-\frac{-V_D}{n\mu C_{ox} W L e^{V_S}}} \cdot 1 + (V_{DD} - V_O)7V_A}{1 + V_O/V_A}.$$  

(1.69)

in which:

$$K_\Delta = e^{\frac{\Delta V_T}{n\mu C_{ox} W L e^{V_S}}} \cdot \frac{1 + \Delta \beta_N/\beta_N}{1 + \Delta \beta_P/\beta_P}.$$  

(1.70)

and

$$\Delta V_T = |\delta V_{T_{0,P}} - \delta V_{T_{0,N}}|.$$  

(1.71)

To study the effect of parameter variation, the term $K_\Delta$ has been added to this equation which includes threshold voltage variation and also variation on transistor $\beta$ value. The nominal value of $K_\Delta$ when there is no parameter variation is one.

Figure 1.12(a) depicts the calculated VTC of an inverter using (1.69) where process variation has been included in the equation. Figures 1.12(b) and 1.12(c) compare the static noise margin and input-output VTC crossover point calculated using (1.69) with the transistor level simulation results. As can be seen, there is a very good agreement between the hand calculations and the transistor level simulation results.

Now, we will try to derive a simplified model for SNM of an inverter operating in subthreshold regime versus process parameters using (1.69). This simplified model can be especially interesting to predict the circuit reliability in course of technology scaling.

From (1.69), it can be seen that the slope of VTC close to the transient point is about:

$$\frac{\partial V_O}{\partial V_I} \approx -\frac{1}{\eta}.$$  

(1.72)

which means the gain of an inverter is mainly limited by the DIBL factor.

To estimate the static noise margin, based on definition, the points in which the slope of VTC becomes -1 should be calculated:

$$\frac{\partial V_O}{\partial V_I} = -1.$$  

(1.73)
The slope of VTC can be calculated using (1.69). Based on this analysis, the static noise margin of an inverter which is biased in subthreshold without including process parameter variations can be estimated by:

$$SNM_0 = \left( \frac{V_{DD}}{2} - U_T \ln \left( \frac{1}{D \cdot (1 - D)} \right) \right) - \frac{1}{\eta} \cdot \left( \frac{V_{DD}}{2} + 2U_T \ln (1 - D) \right)$$  
(1.74)

where:

$$D = \frac{n}{n + 2(1 - \eta)}.$$  
(1.75)

Parameter $D$ depends on subthreshold slope factor, $n$, and DIBL coefficient, $\eta$. As described in (1.74), by increasing of DIBL coefficient, SNM reduces. Therefore, to have a positive SNM value, DIBL coefficient needs to be much smaller than one. It is also noticeable that SNM degrades slightly when $n$, or equivalently subthreshold slope ($S$) increases. Figure 1.13 shows the value of $D$ versus $\eta$. It can also be seen that the estimated value for SNM at (1.74) is very close to the actual value resulted from (1.69).

A very crude approximation for SNM can be:

$$SNM_0 \approx \frac{V_{DD}}{2} \cdot (1 - \eta)$$  
(1.76)

which indicates that SNM reduces almost linearly with increase of $\eta$ value.

**Including process variation:** To derive (1.74), the effect of device parameter variations considered in $K_\Delta$ has been ignored. Including the device variations and after some analysis, it can be shown that SNM will change to:

$$SNM = SNM_0 - \frac{nU_T}{2} \cdot \ln K_\Delta$$  
(1.77)

or

$$SNM = SNM_0 - |\frac{\Delta V_T}{2} + \frac{nU_T}{2} \cdot \ln \left( 1 + \frac{\Delta \beta_N / \beta_N}{1 + \Delta \beta_P / \beta_P} \right)|$$  
(1.78)
It is important to notice that any variation on threshold voltage difference degrades the SNM value regardless of the sign of this variation. Indeed, the maximum SNM can be achieved by setting the crossover point to $V_{DD}/2$ and since variations on threshold voltage difference will move this point to left or right, it degrades the SNM. As the variation on $\beta$, especially logarithm of $\beta$ as appears in (1.78), is negligible in comparison to the variation on threshold voltage,\(^{14}\) although not necessary, this equation can be simplified to

$$SNM = SNM_0 - \frac{\Delta V_T}{2}$$

As crossover point depends on $\Delta V_T = |\delta V_{T0,P} - \delta V_{T0,N}|$, any variation on difference of threshold voltage of PMOS and NMOS devices will be reflected on SNM. In Figure 1.13 the degradation on SNM due to the process variation has been shown. It can be seen that in high DIBL coefficient values, it is really difficult to design a gate with sufficient SNM.

Using (1.79), it is possible to estimate the minimum acceptable size of transistors to have a positive SNM:

$$SNM > 0$$

Using (1.23) and assuming that variation on threshold voltage of PMOS and NMOS devices is uncorrelated and the width of PMOS device is three times larger than NMOS transistors, then the effective width and length of NMOS device\(^{15}\) should be larger than:

$$\sqrt{W_N L_N} > \frac{\sqrt{3} \cdot A_{VT}}{\frac{V_{DD}}{2} - U_T \ln \left(\frac{1}{D(1-D)}\right) - \eta \cdot \left(\frac{V_{DD}}{2} + 2U_T \ln (1-D)\right)} \approx \frac{\sqrt{3} \cdot A_{VT}}{\frac{V_{DD}}{2} (1-\eta)}$$

For this estimation, a coefficient of three has been included in calculation of threshold voltage variation for $3\sigma$ estimation.

\(^{14}\)Based on ITRS suggestion, standard deviation of due to variation on nominal value of $L$ is about 20% [27].

\(^{15}\)Based on this estimation, there is a lower limit on effective physical length and width of transistors. Based on BSIM model, the effective length and width of transistors are [50]:

$$L_e = L + XL - 2 \times dL \approx L + XL - 2 \times DLC$$

$$W_e = W + XW - 2 \times dW \approx W + XW - 2 \times DWC$$
CHAPTER 1. SUBTHRESHOLD MOS FOR ULTRA-LOW POWER

1.4.4 Technology Scaling

This is still an open question that how much technology scaling is desirable or useful for ULP applications. While the main goals of scaling is to reduce the dynamic energy consumption and improve the speed, neither of these two points are a real concern in ULP design. Some preliminary studies show that the technology scaling down to 65nm can bring a considerable advantage for medium throughput applications. However, increase in static energy consumption diminishes this improvement from 65nm to 45nm [14]. Low throughput application suffer more from static power dissipation and hence the improvement due to using more advanced technologies saturates even earlier [14]. Here, we will be trying to provide the necessary basis for answering this type of questions. As will be shown later, with a correct strategy, it is still possible to benefit from technology scaling to improve the system performance with a efficient energy consumption.

Static Noise Margin

To study more precisely the evolution of reliability of the logic cells in course of technology scaling, it is possible to use the analyzes made in Section 1.4.3. Using these analyzes, Figure 1.14(a) shows the static noise margin of an inverter biased in subthreshold regime in different technology nodes. The model parameters for MOS devices have been extracted from [51]. In this figure, the nominal SNM as well as worse case value of SNM including 3σ variations have been shown. As can be seen, for technology nodes close to 45nm, the SNM approaches to zero and hence the circuit reliability becomes a challenge issue. To overcome this problem, either supply voltage should be increased or larger devices must be selected. Figure 1.14(b) shows how much larger the device lengths should be selected to have a nonnegative SNM. For example, in 32nm technology node, the length of each transistor should be selected 1.35 times larger than the minimum effective length to satisfy this requirement (it is assumed that W = 2 × L).

![Figure 1.14](image.png)

**Figure 1.14:** (a) Evolution on static noise margin with technology scaling. The process parameters have been extracted from [51]. This plot shows the supply voltage of the gate which is selected to be equal to the threshold voltage of the devices, nominal SNM value, and also SNM including 3σ variation on device parameters. Here it is assumed that minimum size devices are being used. (b) Ratio that device length should be selected than the minimum size to have a positive SNM (it is assumed that W = 2 × L).

In the next Section, the energy consumption of a digital system operating in subthreshold is
estimated. Based on energy consumption criteria, one can select to increase the length or supply voltage to have the best performance.

**Energy Consumption**

In addition to the speed enhancement, one of the main drivers for technology scaling is implementing more powerful integrated systems with improved energy efficiency. This statement, is generally correct for high speed and high performance applications while in ultra-low power designs using advanced UDSM technologies is still questionable.

To have a rough idea about how efficient can be using of UDSM technologies for ultra-low power designs, some very short analyzes and simulations have been carried out in this Section. As it will be shown in Chapter 3, the power consumption of a CMOS logic circuit with logic depth of \( N \) can be estimated by [52]

\[
P_{\text{diss,CMOS,N}} \approx NI_{\text{leak}}V_{\text{DD}} \sqrt{1 + \frac{\alpha \cdot \zeta}{6} \left( \frac{\gamma^2}{N^2} + \frac{\gamma}{N} - 2 \right)}
\]

(1.84)

where, \( \alpha = f_{\text{op}}/f_{\text{Max}} \) represents the average activity rate of the proposed circuit, \( f_{\text{Max}} = 1/(2t_d) \) is the maximum operation frequency of a single gate, \( \gamma = I_{\text{ON}}/I_{\text{OFF}}, f_{\text{op}} = 1/T, \) and \( \zeta = \lfloor N/2 \rfloor \). As a crude estimation, the maximum operating frequency of a logic gate can be estimated by:

\[
f_{\text{Max}} = \frac{1}{2t_d} \approx \frac{I_{\text{ON}}}{2V_{\text{DD}}C_L}
\]

(1.85)

where \( C_L \) stands for the total capacitive load at the output of a CMOS gate.

Now we can use (1.84) to estimate the power and energy consumption of a digital system in different technology nodes. For this purpose, we use predictive technology model parameters [51] to estimate the power consumption of a system in different CMOS technologies.

**Figure 1.15:** (a) Optimum energy consumption by technology scaling (\( \alpha = 0.5/N, N=20, C_{L0}=50 \text{fF} \)). (b) Supply voltage in which energy consumption can be minimized. This figure also shows the minimum acceptable supply voltage to keep the noise margin positive (\( \alpha = 0.5/N, N=20, C_{L0}=50 \text{fF} \)). (c) Ratio of the optimum supply voltage to device threshold voltage by technology scaling (\( \alpha = 0.5/N, N=20, C_{L0}=50 \text{fF} \)). \( C_{L0} \) is the average interconnect capacitive loading on each gate. The total capacitive load is sum of this capacitance and the input capacitance of the following CMOS gates.

Now we can use (1.84) to estimate the power and energy consumption of a digital system in different technology nodes. For this purpose, we use predictive technology model parameters [51] to estimate the power consumption of a system in different CMOS technologies.
As an example, assume that the average system logic depth is $N=20$, the activity rate is $\alpha = 1/(2N)$, and the load capacitance is mainly dominated by the interconnections with an average loading effect of about $C_L=50\text{fF}$. A small fanout of two has been considered for each gate, as well. To have a fair estimation, "low power" device option with higher threshold voltage and less gate leakage current has been selected for this calculations.

The results for this estimation is shown in Figure 1.15. In Figure 1.15(a), the minimum achievable energy consumption is shown with the black line (with ◦ sign). In each technology node, the supply voltage has been swept in order to find the minimum energy consumption. The corresponding supply voltage for minimum energy consumption is shown in Figure 1.15(b) while Figure 1.15(c) shows this supply voltage normalized to the device threshold voltage at the corresponding technology node. As can be seen in Figure 1.15(a), by scaling the technology from 0.25\(\mu\)m to around 90nm, the energy consumption can be reduced. However, as technology continues to scale down, the minimum possible energy consumption increases. In other words, technology scaling below 90nm does not help to reduce the energy consumption of the ultra-low power circuits with the aforementioned conditions for activity rate and load capacitance.

It is important to notice that as depicted in Figure 1.15(c), for optimized energy consumption, the supply voltage needs be selected more and more close to the threshold voltage when the device feature sizes are decreased. This is mainly due to the leakage current enhancement in more advanced technologies.

To have a more precise answer for this question that if technology scaling is useful for ULP, we should also consider the process variation. In practice it is not possible to achieve to this level of energy consumption mainly because in such a low supply voltage the static noise margin of the logic cells will be unacceptably low. As illustrated in Figure 1.15(b), the supply voltage for minimizing energy consumption is well below the acceptable level of $V_{DD}$ to have a positive SNM (shown with the blue line). This means that either the supply voltage or the device sizes need to be increase to improve the SNM value to an acceptable level.

There is no clear guideline to how choose the supply voltage and device sizes to have the minimum energy consumption penalty. In Figure 1.15(a) the results for the two possibilities, i.e. supply voltage enhancement or up-sizing the devices, are compared. The blue line shows the energy consumption when the supply voltage is set to the minimum $V_{DD}$ value giving nonnegative SNM. As can be seen, using this approach the energy consumption increases considerably. However, technology scaling again helps to reduce the energy consumption until about 65nm. For smaller technology nodes, the energy consumption increases and hence there is no benefit to use such an advance technologies for ultra-low power applications.

Now lets study the results for up-sizing the devices without increasing the supply voltage from its optimum value. Increasing the size of transistors helps to reduce the device parameter variations and hence there is no need to scale up the supply voltage. Here, in each technology node, the size of PMOS and NMOS devices are scaled up based on (1.83) such that SNM value becomes positive. As can be seen, for technologies above 0.18\(\mu\)m, the energy penalty in this case is much less than the case that supply voltage has been increased. This result is valid for very advanced technologies such as 45nm and 32nm as well. For technologies between 65nm to 0.18\(\mu\)m, the energy consumption of scaled supply voltage system will be lower.

Of course the result for the comparison can be changed by changing the system specifications such as activity rate or loading effect. In any case, the relationships derived in this Section can give a clear insight about the main design tradeoffs for implementing ultra-low power systems in advanced CMOS technologies. For example, Figures 1.16(a)-(c) show the same graphs for a different condition in which...
Figure 1.16: (a) Optimum energy consumption by technology scaling ($\alpha = 0.1/N$, $N=20$, $C_{L0}=5\text{fF}$). (b) Supply voltage in which energy consumption can be minimized. This figure also shows the minimum acceptable supply voltage to keep the noise margin positive ($\alpha = 0.1/N$, $N=20$, $C_{L0}=5\text{fF}$). (c) Ratio of the optimum supply voltage to device threshold voltage by technology scaling ($\alpha = 0.1/N$, $N=20$, $C_{L0}=5\text{fF}$). $C_{L0}$ is the average interconnect capacitive loading on each gate. The total capacitive load is sum of this capacitance and the input capacitance of the following CMOS gates.

activity rate is very low while load capacitance is mainly dominated by the gate capacitance instead of interconnections. In this case, $V_{DD}$ scaling is more efficient than device up-sizing for all the technology nodes. Here, since the activity rate is lower, the leakage current is more pronounced and hence the optimum $V_{DD}$ value shown in Figure 1.16(b) is higher than the previous case (Figure 1.15(b)).

As a conclusion, a very careful design strategy for selecting optimum supply voltage or choosing proper device sizes is required to benefit from technology scaling. Even in very deep technology nodes, still there is this possibility to reduce the energy consumption. Of course this statement depends highly on high level system specifications such as logic depth, activity rate, interconnections, and etc.

In the rest of this report, some techniques for implementing ULP digital and analog circuits based on subthreshold MOS devices will be described. The emphasize here is to address the main existing design issues such as leakage (static) current reduction and implementing reliable circuits in very low current densities.
Bibliography


[51] Predictive Technology Model, [online], http://www.eas.asu.edu/ptm/.