Chapter 13

Nonlinearity and Mismatch

In Chapters 6 and 7, we dealt with two types of nonidealities, namely, frequency response and noise, that limit the performance of analog circuits. In this chapter, we study two other imperfections that prove critical in high-precision analog design and trade with many other performance parameters. These effects are nonlinearity and mismatch.

We first define metrics for quantifying the effects of nonlinearity. Next, we study nonlinearity in differential circuits and feedback systems and examine several linearization techniques. We then deal with the problem of mismatch and dc offsets in differential circuits. Finally, we consider a number of offset cancellation methods and describe the effect of offset cancellation on random noise.

13.1 Nonlinearity

13.1.1 General Considerations

As we have observed in the large-signal analysis of single-stage and differential amplifiers, circuits usually exhibit a nonlinear input/output characteristic. Depicted in Fig. 13.1, such a characteristic deviates from a straight line as the input swing increases. Two examples

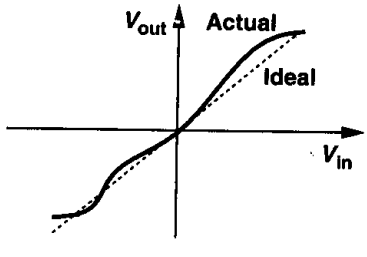


Figure 13.1 Input/output characteristic of a nonlinear system.

are shown in Fig. 13.2. In a common-source stage or a differential pair, the output variation becomes heavily nonlinear as the input level increases. In other words, for a small input swing, the output is a reasonable replica of the input but for large swings the output exhibits "saturated" levels.

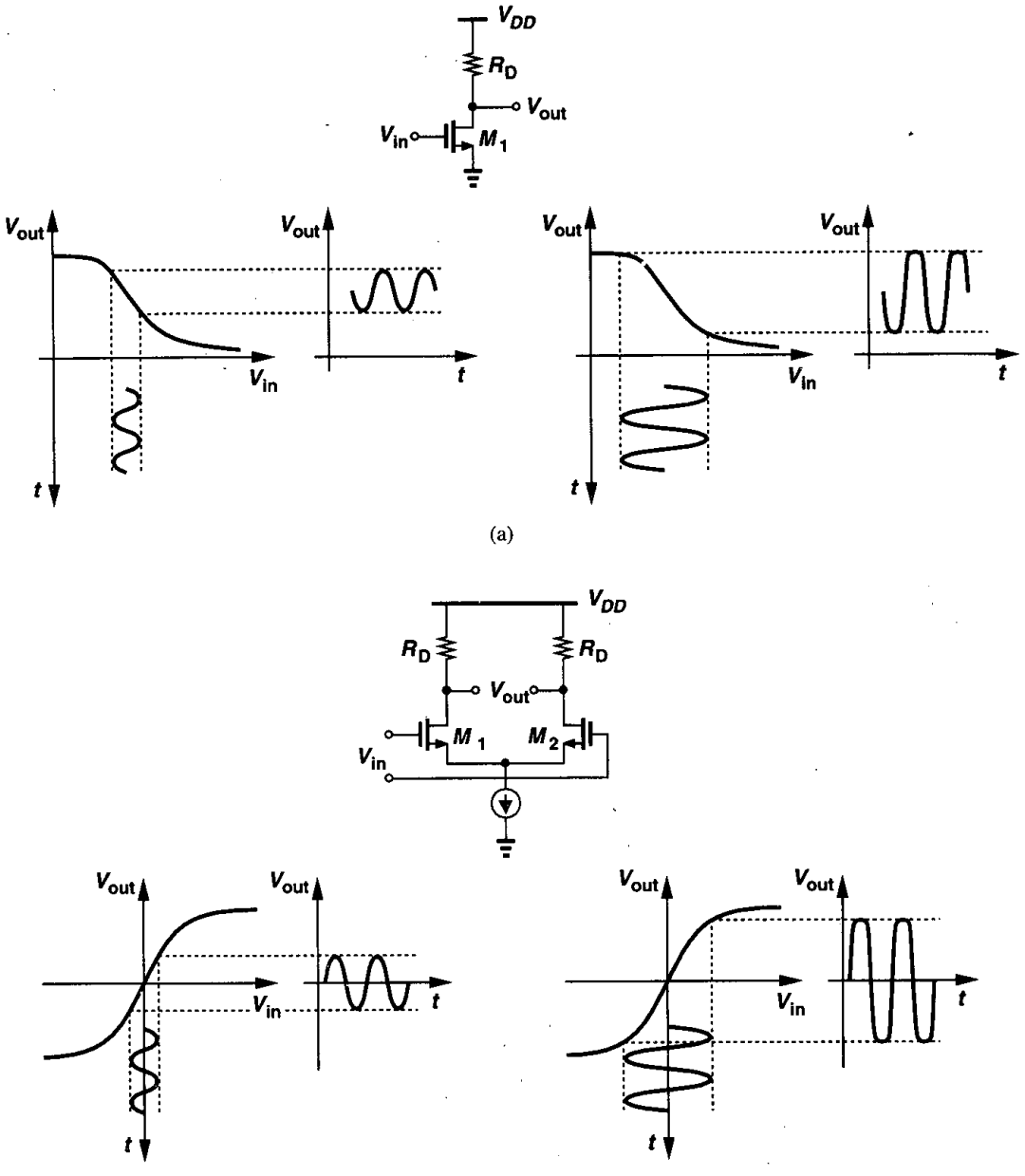


Figure 13.2 Distortion in (a) a common-source stage and (b) a differential pair.

The nonlinear behavior of a circuit can also be viewed as *variation* of the slope and hence the small-signal gain with the input level. Illustrated in Fig. 13.3, this observation means that a given incremental change at the input results in different incremental changes at the output depending on the input dc level.

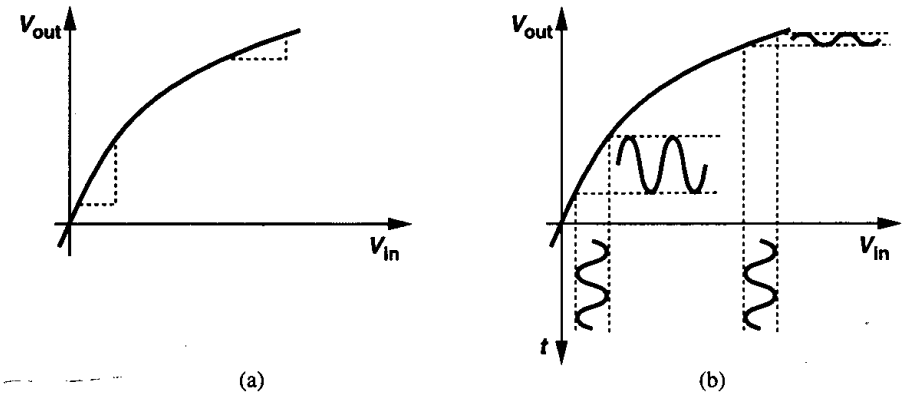


Figure 13.3 Variation of small-signal gain in a nonlinear amplifier.

In many analog circuits, precision requirements mandate relatively small nonlinearities, making it possible to approximate the input/output characteristic by a Taylor expansion in the range of interest:

$$y(t) = \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t) + \cdots. \tag{13.1}$$

For small x, $y(t) \approx \alpha_1 x$, indicating that α_1 is the small-signal gain in the vicinity of $x \approx 0$.

How is the nonlinearity quantified? A simple method is to identify α_1, α_2, etc., in (13.1). Another metric that proves useful in practice is to specify the maximum deviation of the characteristic from an ideal one (i.e., a straight line). As shown in Fig. 13.4, for the voltage range of interest, $[0 \quad V_{in,max}]$, we pass a straight line through the end points of the actual characteristic, obtain the maximum deviation, ΔV, and normalize the result to the maximum output swing, $V_{out,max}$. For example, we say an amplifier exhibits 1% nonlinearity ($\Delta V / V_{out,max} = 0.01$) for an input range of 1 V.

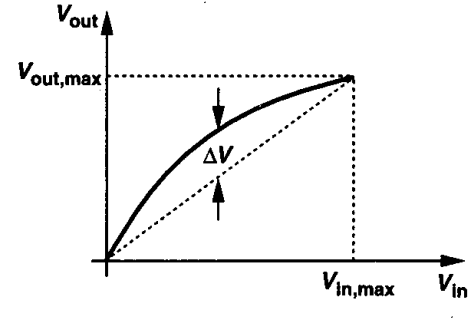


Figure 13.4 Definition of nonlinearity.

Example 13.1

The input/output characteristic of a differential amplifier is approximated as $y(t) = \alpha_1 x(t) + \alpha_3 x^3(t)$. Calculate the maximum nonlinearity if the input range is from $x = -x_{max}$ to $x = +x_{max}$.

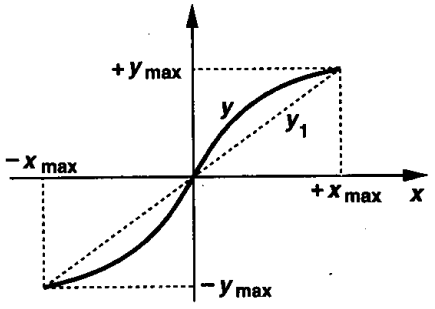


Figure 13.5

Solution

As depicted in Fig. 13.5, we can express the straight line passing through the end points as

$$y_1 = \frac{\alpha_1 x_{max} + \alpha_3 x_{max}^3}{x_{max}} x \tag{13.2}$$

$$= \left(\alpha_1 + \alpha_3 x_{max}^2\right) x. \tag{13.3}$$

The difference between y and y_1 is therefore equal to

$$\Delta y = y - y_1 \tag{13.4}$$

$$= \alpha_1 x + \alpha_3 x^3 - \left(\alpha_1 + \alpha_3 x_{max}^2\right) x. \tag{13.5}$$

Setting the derivative of Δy with respect to x to zero, we have $x = x_{max}/\sqrt{3}$ and the maximum deviation is equal to $2\alpha_3 x_{max}^3/(3\sqrt{3})$. Normalized to the maximum output, the nonlinearity is obtained as

$$\frac{\Delta y}{y_{max}} = \frac{2\alpha_3 x_{max}^3}{3\sqrt{3} \times 2\left(\alpha_1 x_{max} + \alpha_3 x_{max}^3\right)}. \tag{13.6}$$

Note that the factor of 2 in the denominator is included because the maximum peak-to-peak output swing is equal to $2(\alpha_1 x_{max} + \alpha_3 x_{max}^3)$. For small nonlinearities, we can neglect $\alpha_3 x_{max}^3$ with respect to $\alpha_1 x_{max}$, arriving at

$$\frac{\Delta y}{y_{max}} \approx \frac{\alpha_3}{3\sqrt{3}\alpha_1} x_{max}^2. \tag{13.7}$$

Note that the relative nonlinearity is proportional to the square of the maximum input swing in this example.

The nonlinearity of a circuit can also be characterized by applying a sinusoid at the input and measuring the harmonic content of the output. Specifically, if in (13.1), $x(t) = A\cos\omega t$, then

$$y(t) = \alpha_1 A\cos\omega t + \alpha_2 A^2\cos^2\omega t + \alpha_3\cos^3\omega t + \cdots \tag{13.8}$$

$$= \alpha_1 A\cos\omega t + \frac{\alpha_2 A^2}{2}[1+\cos(2\omega t)] + \frac{\alpha_3 A^3}{4}[3\cos\omega t + \cos(3\omega t)] + \cdots. \tag{13.9}$$

We observe that higher-order terms yield higher harmonics. In particular, even-order terms and odd-order terms result in even and odd harmonics, respectively. Note that the magnitude of the nth harmonic grows roughly in proportion to the nth power of the input amplitude. Called "harmonic distortion," this effect is usually quantified by summing the power of all of the harmonics (except that of the fundamental) and normalizing the result to the power of the fundamental. Such a metric is called the "total harmonic distortion" (THD). For a third-order nonlinearity:

$$THD = \frac{(\alpha_2 A^2/2)^2 + (\alpha_3 A^3/4)^2}{(\alpha_1 A + 3\alpha_3 A^3/4)^2}. \tag{13.10}$$

Harmonic distortion is undesirable in most signal processing applications, including audio and video systems. High-quality audio products such as compact disc (CD) players require a THD of about 0.01% (−80 dB) and video products, about 0.1% (−60 dB).

13.1.2 Nonlinearity of Differential Circuits

Differential circuits exhibit an "odd-symmetric" input/output characteristic, i.e., $f(-x) = -f(x)$. For the Taylor expansion of (13.1) to be an odd function, all of the even-order terms, α_{2j} must be zero:

$$y(t) = \alpha_1 x(t) + \alpha_3 x^3(t) + \alpha_5 x^5(t) + \cdots, \tag{13.11}$$

indicating that a differential circuit driven by a differential signal produces no even harmonics. This is another very important property of differential operation.

In order to appreciate the reduction of nonlinearity obtained by differential operation, let us consider the two amplifiers shown in Fig. 13.6, each of which is designed to provide a small-signal voltage gain of

$$|A_v| \approx g_m R_D \tag{13.12}$$

$$= \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})R_D. \tag{13.13}$$

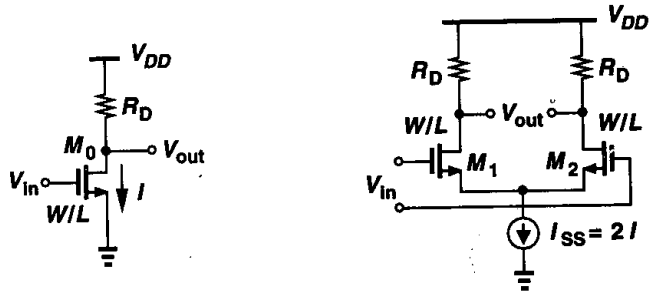


Figure 13.6 Single-ended and differential amplifiers providing the same voltage gain.

Suppose a signal $V_m \cos \omega t$ is applied to each circuit. Examining only the drain currents for simplicity, we can write for the common-source stage:

$$\begin{aligned} I_{D0} &= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH} + V_m \cos\omega t)^2 \\ &= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 + \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})V_m \cos\omega t \\ &\quad + \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_m^2 \cos^2\omega t \\ &= I + \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})V_m \cos\omega t + \frac{1}{4}\mu_n C_{ox}\frac{W}{L}V_m^2[1 + \cos(2\omega t)]. \end{aligned} \tag{13.14}$$

Thus, the amplitude of the second harmonic, A_{HD2}, normalized to that of the fundamental, A_F, is

$$\frac{A_{HD2}}{A_F} = \frac{V_m}{4(V_{GS} - V_{TH})}. \tag{13.15}$$

On the other hand, for M_1 and M_2 in Fig. 13.6, we have from Chapter 4:

$$I_{D1} - I_{D2} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{in}\sqrt{\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}} - V_{in}^2} \tag{13.16}$$

$$= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{in}\sqrt{4(V_{GS} - V_{TH})^2 - V_{in}^2}. \tag{13.17}$$

If $|V_{in}| \ll V_{GS} - V_{TH}$, then

$$I_{D1} - I_{D2} = \mu_n C_{ox} \frac{W}{L} V_{in}(V_{GS} - V_{TH}) \sqrt{1 - \frac{V_{in}^2}{4(V_{GS} - V_{TH})^2}} \tag{13.18}$$

$$\approx \mu_n C_{ox} \frac{W}{L} V_{in}(V_{GS} - V_{TH}) \left[1 - \frac{V_{in}^2}{8(V_{GS} - V_{TH})^2}\right] \tag{13.19}$$

$$= \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH}) \left[V_m \cos \omega t - \frac{V_m^3 \cos^3 \omega t}{8(V_{GS} - V_{TH})^2}\right]. \tag{13.20}$$

Since $\cos^3 \omega t = [3 \cos \omega t + \cos(3\omega t)]/4$, we obtain,

$$I_{D1} - I_{D2} = g_m \left[V_m - \frac{3V_m^3}{32(V_{GS} - V_{TH})^2}\right] \cos \omega t - g_m \frac{V_m^3 \cos(3\omega t)}{32(V_{GS} - V_{TH})^2}. \tag{13.21}$$

If $V_m \gg 3V_m^3/[8(V_{GS} - V_{TH})^2]$, then

$$\frac{A_{HD3}}{A_F} \approx \frac{V_m^2}{32(V_{GS} - V_{TH})^2}. \tag{13.22}$$

Comparison of (13.15) and (13.22) indicates that the differential circuit exhibits much less distortion than its single-ended counterpart while providing the same voltage gain and output swing. For example, if $V_m = 0.2(V_{GS} - V_{TH})$, (13.15) and (13.22) yield a distortion of 5% and 0.125%, respectively.

While achieving a lower distortion, the differential pair consumes twice as much power as the CS stage because $I_{SS} = 2I$. The key point, however, is that even if the bias current of M_0 is raised to $2I$, (13.15) predicts that the distortion decreases by only a factor of $\sqrt{2}$ (with W/L maintained constant).

13.1.3 Effect of Negative Feedback on Nonlinearity

In Chapter 8, we observed that negative feedback makes the closed-loop gain relatively independent of the op amp's open-loop gain. Since nonlinearity can be viewed as variation of the small-signal gain with the input level, we expect that negative feedback suppresses this variation as well, yielding higher linearity for the closed-loop system.

Analysis of nonlinearity in a feedback system is quite complex. Here, we consider a simple, "mildly nonlinear" system to gain more insight. The reason is that, if properly designed, a feedback amplifier exhibits only small distortion components, lending itself to this type of analysis.

Let us assume that the core amplifier in the system of Fig. 13.7 has an input-output characteristic $y \approx \alpha_1 x + \alpha_2 x^2$. We apply a sinusoidal input $x(t) = V_m \cos \omega t$, postulating that the output contains a fundamental component and a second harmonic and hence can

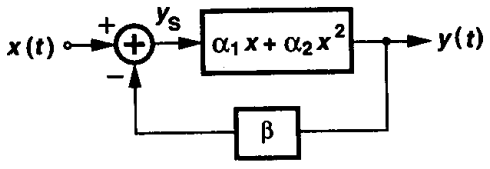


Figure 13.7 Feedback system incorporating a nonlinear feedforward amplifier.

be approximated as $y \approx a\cos\omega t + b\cos 2\omega t$.[1] Our objective is to determine a and b. The output of the subtractor can be written as

$$y_S = x(t) - \beta y(t) \tag{13.23}$$

$$= V_m \cos\omega t - \beta(a\cos\omega t + b\cos 2\omega t) \tag{13.24}$$

$$= (V_m - \beta a)\cos\omega t - \beta b\cos 2\omega t. \tag{13.25}$$

This signal experiences the nonlinearity of the feedforward amplifier, thereby producing an output given by:

$$y(t) = \alpha_1[(V_m - \beta)\cos\omega t - \beta b\cos 2\omega t] + \alpha_2[(V_m - \beta a)\cos\omega t - \beta b\cos 2\omega t]^2 \tag{13.26}$$

$$= [\alpha_1(V_m - \beta a) - \alpha_2(V_m - \beta a)\beta b]\cos\omega t + \left[-\alpha_1\beta b + \frac{\alpha_2(V_m - \beta a)^2}{2}\right]\cos 2\omega t + \cdots. \tag{13.27}$$

The coefficients of $\cos\omega t$ and $\cos 2\omega t$ in (13.27) must be equal to a and b, respectively:

$$a = (\alpha_1 - \alpha_2\beta b)(V_m - \beta a) \tag{13.28}$$

$$b = -\alpha_1\beta b + \frac{\alpha_2(V_m - \beta a)^2}{2}. \tag{13.29}$$

The assumption of small nonlinearity implies that both α_2 and b are small quantities, yielding $a \approx \alpha_1(V_m - \beta a)$ and hence

$$a = \frac{\alpha_1}{1 + \beta\alpha_1} V_m, \tag{13.30}$$

which is to be expected because $\beta\alpha_1$ is the loop gain. To calculate b, we write

$$V_m - \beta a \approx \frac{a}{\alpha_1}, \tag{13.31}$$

[1] Note that higher harmonics and phase shifts through the system are neglected.

thus expressing (13.29) as

$$b = -\alpha_1 \beta b + \frac{1}{2}\alpha_2 \left(\frac{a}{\alpha_1}\right)^2. \tag{13.32}$$

That is,

$$b(1 + \alpha_1 \beta) = \frac{\alpha_2}{2}\left(\frac{a}{\alpha_1}\right)^2 \tag{13.33}$$

$$= \frac{\alpha_2}{2\alpha_1^2}\frac{\alpha_1^2}{(1 + \beta\alpha_1)^2} V_m^2. \tag{13.34}$$

It follows that

$$b = \frac{\alpha_2 V_m^2}{2}\frac{1}{(1 + \beta\alpha_1)^3}. \tag{13.35}$$

For a meaningful comparison, we normalize the amplitude of the second harmonic to that of the fundamental:

$$\frac{b}{a} = \frac{\alpha_2 V_m}{2}\frac{1}{\alpha_1}\frac{1}{(1 + \beta\alpha_1)^2}. \tag{13.36}$$

Without feedback, on the other hand, such a ratio would be equal to $(\alpha_2 V_m^2/2)/\alpha_1 V_m = \alpha_2 V_m/(2\alpha_1)$. Thus, the relative magnitude of the second harmonic has dropped by a factor of $(1 + \beta\alpha_1)^2$.

As described in Chapter 8, a feedback circuit employing a feedforward amplifier with a finite gain suffers from gain error. For a feedforward gain of A_0 and a feedback factor of β, the relative gain error is approximately equal to $1/(\beta A_0)$. If the feedforward amplifier exhibits nonlinearity, it is possible to derive a simple relationship between the gain error and maximum nonlinearity of the overall feedback circuit. As illustrated in Fig. 13.8, we draw two straight lines, one representing the ideal characteristic (with a slope $1/\beta$) and another passing through the end points of the actual characteristic. We note that with this construction, the nonlinearity, Δy_2, is always smaller than the gain error, Δy_1. This is of course true only if the small-signal gain drops monotonically as x goes from 0 to x_{max}, a

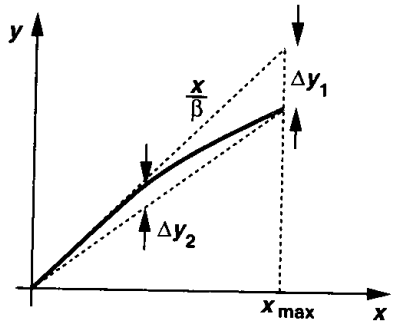


Figure 13.8 Gain error and nonlinearity in a feedback system.

typical behavior in most analog circuits. Thus, a sufficient condition to ensure $\Delta y_2 < \epsilon$ is to guarantee that $\Delta y_1 < \epsilon$ by choosing a high open-loop gain for the amplifier.

The above condition is often applied in analog design because it is much easier to predict the open-loop gain than its nonlinearity. Of course, this simplification is obtained at the cost of a pessimistic choice of the amplifier's gain, an issue that becomes more serious as short-channel devices limit the voltage gain that can be achieved.

13.1.4 Capacitor Nonlinearity

In switched-capacitor circuits, the voltage dependence of capacitors may introduce substantial distortion. While for a linear capacitor we have $Q = CV$, for a voltage-dependent capacitor we must write $dQ = C\,dV$. Thus, the total charge on a capacitor sustaining a voltage V_1 is

$$Q(V_1) = \int_0^{V_1} C\,dV. \tag{13.37}$$

To study the effect of capacitor nonlinearity, we express each capacitor as $C = C_0(1 + \alpha_1 V + \alpha_2 V^2 + \cdots)$.

Let us consider the noninverting amplifier of Fig. 12.41(a), repeated in Fig. 13.9, as an example. At the beginning of the amplification mode, C_1 has a voltage equal to V_{in0} and C_2 a voltage of zero. Assuming $C_1 \approx MC_0(1 + \alpha_1 V)$, where M is the nominal closed-loop gain ($C_1 = MC_2$), we obtain the charge across C_1 as

$$Q_1 = \int_0^{V_{in0}} C_1\,dV \tag{13.38}$$

$$= \int_0^{V_{in0}} MC_0(1 + \alpha_1 V)\,dV \tag{13.39}$$

$$= MC_0 V_{in0} + MC_0 \frac{\alpha_1}{2} V^2. \tag{13.40}$$

Similarly, if $C_2 \approx C_0(1 + \alpha_1 V)$, then the charge on this capacitor at the end of the amplification mode is

$$Q_2 = \int_0^{V_{out}} C_2\,dV \tag{13.41}$$

$$= C_0 V_{out} + C_0 \frac{\alpha_1}{2} V_{out}^2. \tag{13.42}$$

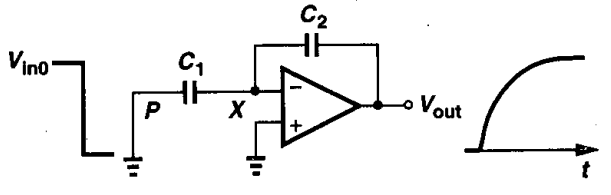


Figure 13.9 Effect of capacitor nonlinearity.

Equating Q_1 and Q_2 and solving for V_{out}, we have

$$V_{out} = \frac{1}{\alpha_1}\left(-1 + \sqrt{1 + M\alpha_1^2 V_{in0}^2 + 2M\alpha_1 V_{in0}}\right). \tag{13.43}$$

The last two terms under the square root are usually much less than unity and, since for $\epsilon \ll 1$, $\sqrt{1+\epsilon} \approx 1 + \epsilon/2 - \epsilon^2/8$, we can write

$$V_{out} \approx MV_{in0} + (1 - M)\frac{M\alpha_1}{2} V_{in0}^2. \tag{13.44}$$

The second term in the above equation represents the nonlinearity resulting from the voltage dependence of the capacitor.

13.1.5 Linearization Techniques

While amplifiers using "global" feedback (e.g., the switched-capacitor topologies of Chapter 12) can achieve a high linearity, stability and settling issues of feedback circuits limit their usage in high-speed applications. For this reason, many other techniques have been invented to linearize amplifiers with less compromise in speed.

The principle behind linearization is to reduce the dependence of the gain of the circuit upon the input level. This usually translates into making the gain relatively independent of the transistor bias currents.

The simplest linearization method is source degeneration by means of a linear resistor. As shown in Fig. 13.10 for a common-source stage and revealed by the observations in the

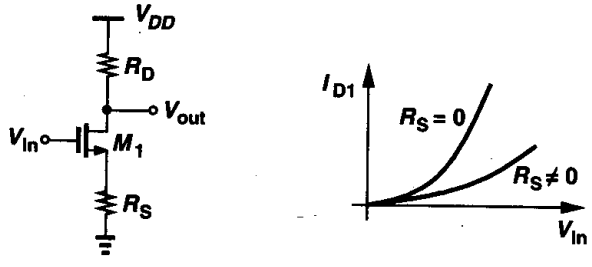


Figure 13.10 Common-source stage with resistive degeneration.

previous section, degeneration reduces the signal swing applied between the gate and the source of the transistor, thereby making the input/output characteristic more linear. From another point of view, neglecting body effect, we can write the overall transconductance of the stage as

$$G_m = \frac{g_m}{1 + g_m R_S}, \tag{13.45}$$

which for large $g_m R_S$ approaches $1/R_S$, an input-independent value.

Note that the amount of linearization depends on $g_m R_S$ rather on R_S alone. With a relatively constant G_m, the voltage gain, $G_m R_D$, is also relatively independent of the input and the amplifier is linearized.

Example 13.2

A common-source stage biased at a current I_1 experiences an input voltage swing that varies the drain current from $0.75I_1$ to $1.25I_1$. Calculate the variation of the small-signal voltage gain (a) with no degeneration, (b) with degeneration such that $g_m R_S = 2$, where g_m denotes the transconductance at $I_D = I_1$.

Solution

Assuming square-law behavior, we have $g_m \propto \sqrt{I_D}$. For the case of no degeneration:

$$\frac{g_{m,high}}{g_{m,low}} = \sqrt{\frac{1.25}{0.75}}. \tag{13.46}$$

With $g_m R_S = 2$,

$$\frac{G_{m,high}}{G_{m,low}} = \frac{\dfrac{\sqrt{1.25}g_m}{1+\sqrt{1.25}g_m R_S}}{\dfrac{\sqrt{0.75}g_m}{1+\sqrt{0.75}g_m R_S}} \tag{13.47}$$

$$= \sqrt{\frac{1.25}{0.75}} \cdot \frac{1+2\sqrt{0.75}}{1+2\sqrt{1.25}} \tag{13.48}$$

$$= 0.84\sqrt{\frac{1.25}{0.75}}. \tag{13.49}$$

Thus, degeneration decreases the variation of the small-signal gain by approximately 16% in this case.

Resistive degeneration presents trade-offs between linearity, noise, power dissipation, and gain. For reasonable input voltage swings (e.g., 1 V_{pp}), it may be quite difficult to achieve even a voltage gain of 2 in a common-source stage if the nonlinearity is to remain below 1%.

A differential pair can be degenerated as shown in Figs. 13.11(a) and (b). In Fig. 13.11(a), I_{SS} flows through the degeneration resistors, thereby consuming a voltage headroom of $I_{SS}R_S/2$, an important issue if a high level of degeneration is required. The circuit of Fig. 13.11(b), on the other hand, does not involve this issue but it suffers from a slightly higher noise (and offset voltage) because the two tail current sources introduce some differential error. The reader can prove that if the output noise current of each current source is equal to $\overline{I_n^2}$, then the input-referred noise voltage of the circuit of Fig. 13.11(b) is higher than that of Fig. 13.11(a) by $2\overline{I_n^2}R_S^2$.

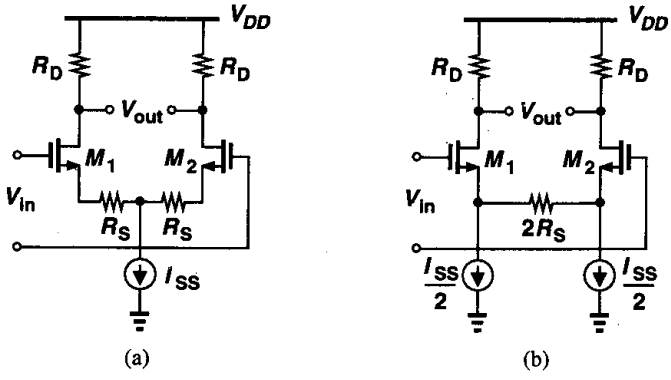


Figure 13.11 Source degeneration applied to a differential pair.

Resistive degeneration requires high-quality resistors, a commodity unavailable in many of today's CMOS technologies (Chapter 17). As depicted in Fig. 13.12, the resistor can be replaced by a MOSFET operating in deep triode region. However, for large input swings, M_3 may not remain in deep triode region, thereby experiencing substantial change in its on-resistance. Furthermore, V_b must track the input common-mode level so that R_{on3} can be defined accurately.

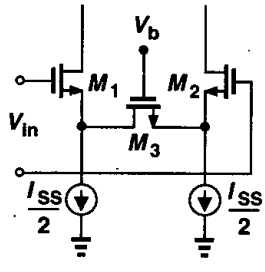


Figure 13.12 Differential pair degenerated by a MOSFET operating in deep triode region.

Another linearization technique is illustrated in Fig. 13.13 [1]. Here, M_3 and M_4 are in deep triode region if $V_{in} = 0$. As the gate voltage of M_1 becomes more positive than the gate voltage of M_2, transistor M_3 stays in the triode region because $V_{D3} = V_{G3} - V_{GS1}$ whereas

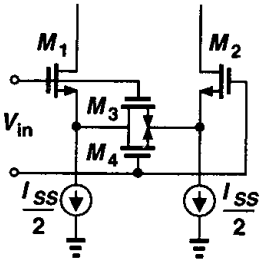


Figure 13.13 Differential pair degenerated by two MOSFETs operating in the triode region.

M_4 eventually enters the saturation region because its drain voltage rises and its gate and source voltages fall. Thus, the circuit remains relatively linear even if one degeneration device goes into saturation. For the widest linear region, [1] suggests $(W/L)_{1,2} \approx 7(W/L)_{3,4}$.

A linearization technique avoiding the use of resistors is based on the observation that a MOSFET operating in the triode region can provide a linear I_D/V_{GS} characteristic if its drain-source voltage is held constant: $I_D = (1/2)\mu C_{ox}(W/L)[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2]$. Illustrated in Fig. 13.14, the technique employs amplifiers A_1 and A_2 along with cascode devices M_3 and M_4 to force V_X and V_Y to be equal to V_b for varying input levels.

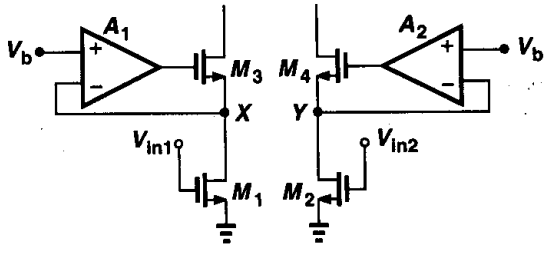


Figure 13.14 Differential pair using input devices operating in the triode region.

This circuit suffers from several drawbacks. First, the transconductance of M_1 and M_2, equal to $\mu_n C_{ox}(W/L)V_{DS}$, is relatively small because V_{DS} must be low enough to ensure each input transistor remains in the triode region. Second, the input common-mode level must be tightly controlled and it must track V_b so as to define I_{D1} and I_{D2}. Third, M_3, M_4, and the two auxiliary amplifiers contribute substantial noise to the output.

Another approach to linearizing voltage amplifiers is to perform "post-correction." Illustrated in Fig. 13.15, the idea is to view the amplifier as a voltage-to-current (V/I) converter followed by a current-to-voltage (I/V) converter. If the V/I converter can be described as $I_{out} = f(V_{in})$ and the I/V converter as $V_{out} = f^{-1}(I_{in})$, then V_{out} is a linear function of V_{in}. That is, the second stage corrects the nonlinearity introduced by the first stage. As an example, recall from Chapter 4 that for the circuit shown in Fig. 13.16(a), we have

$$V_{in1} - V_{in2} = V_{GS1} - V_{GS2} \tag{13.50}$$

$$= \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{1,2}}} - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{1,2}}}. \tag{13.51}$$

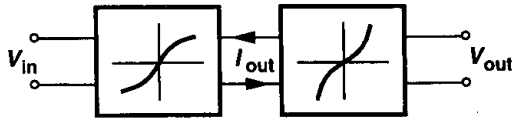


Figure 13.15 Voltage amplifier viewed as a cascade of two nonlinear stages.

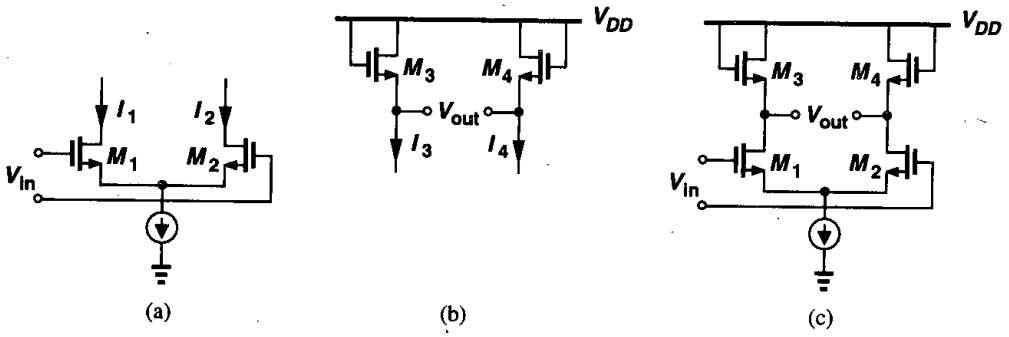


Figure 13.16 (a) Differential pair with nonlinear I/V characteristic, (b) diode-connected devices with nonlinear V/I characteristic, (c) circuit having linear input/output characteristic.

We also note that for the circuit shown in Fig. 13.16(b),

$$V_{out} = V_{GS3} - V_{GS4} \tag{13.52}$$

$$= \sqrt{\frac{2I_3}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4}}} - \sqrt{\frac{2I_4}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4}}}, \tag{13.53}$$

where channel-length modulation and body effect are neglected. It follows that for the circuit shown in Fig. 13.16(c),

$$V_{out} = \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4}}} - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4}}} \tag{13.54}$$

$$= \frac{1}{\sqrt{\left(\frac{W}{L}\right)_{3,4}}}(V_{in1} - V_{in2})\, sqrt\left(\frac{W}{L}\right)_{1,2}. \tag{13.55}$$

Thus, as derived in Chapter 4, the voltage gain is equal to

$$A_v = \sqrt{\frac{\left(\frac{W}{L}\right)_{1,2}}{\left(\frac{W}{L}\right)_{3,4}}}, \tag{13.56}$$

a quantity independent of the bias currents of the transistors.

In practice, body effect and other nonidealities in short-channel devices give rise to nonlinearity in this circuit. Furthermore, as the differential input level increases, driving M_1

or M_2 into the subthreshold region, Eqs. (13.51) and (13.53) no longer hold and the gain drops sharply.

13.2 Mismatch

Our study of amplifiers in the previous chapters has mostly assumed that the circuits are perfectly symmetric, i.e., the two sides exhibit identical properties and bias currents. In reality, however, nominally-identical devices suffer from a finite mismatch due to uncertainties in each step of the manufacturing process. For example, as illustrated in Fig. 13.17, the gate dimensions of MOSFETs suffer from random, microscopic variations and hence mismatches between the equivalent lengths and widths of two transistors that are identically laid out. Also, MOS devices exhibit threshold voltage mismatch because, from (2.1), V_{TH} is a function of the doping levels in the channel and the gate, and these levels vary randomly from one device to another.

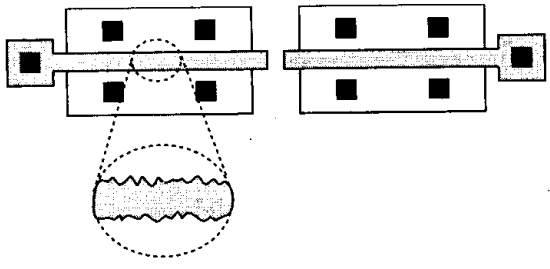


Figure 13.17 Random mismatches due to microscopic variations in device dimensions.

Study of mismatch consists of two steps: (1) identify and formulate the mechanisms that lead to mismatch between devices; (2) analyze the effect of device mismatches upon the performance of circuits. Unfortunately, the first step is quite complex and heavily dependent on the fabrication technology and the layout, often requiring actual measurements of mismatches. For example, the achievable mismatch between capacitors is typically quoted to be 0.1%, but this value is not derived from any fundamental quantities. We therefore consider only some basic trends and intuitive results. Layout techniques for minimum mismatch are described in Chapter 18.

Expressing the characteristics of a MOSFET in saturation as $I_D = (1/2)\mu C_{ox}(W/L)(V_{GS} - V_{TH})^2$, we observe that mismatches between μ, C_{ox}, W, L, and V_{TH} result in mismatches between drain currents (for a given V_{GS}) or gate-source voltages (for a given drain current) of two nominally-identical transistors. Intuitively, we expect that as W and L increase, their relative mismatches, $\Delta W/W$ and $\Delta L/L$, respectively, decrease, i.e., larger devices exhibit smaller mismatches. A more important observation is that all of the mismatches decrease as the *area* of the transistor, WL, increases. For example, increasing W reduces both $\Delta W/W$ *and* $\Delta L/L$. This is because as WL increases, random variations experience greater "averaging," thereby falling in magnitude. For the case depicted in Fig. 13.18, $\Delta L_2 < \Delta L_1$ because, if the device is viewed as many small parallel transistors (Fig. 13.19), each having a width W_0, then we can write the equivalent length as

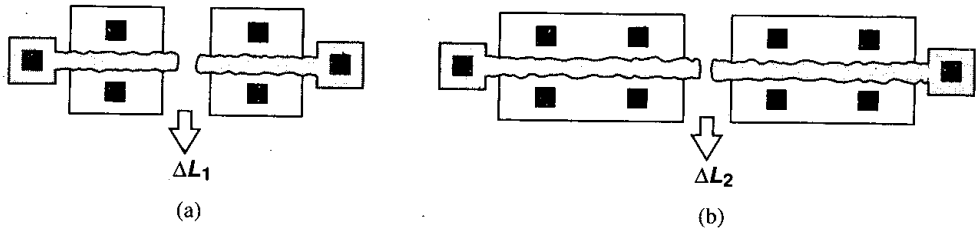


Figure 13.18 Reduction of length mismatch as a result of increasing the width.

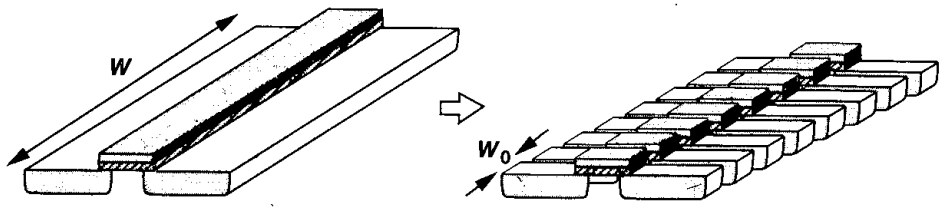


Figure 13.19 Wide MOSFET viewed as a parallel combination of narrow devices.

$L_{eq} \approx (L_1 + L_2 + \cdots + L_n)/n$. The overall variation is therefore given by

$$\Delta L_{eq} \approx \left(\Delta L_1^2 + \Delta L_2^2 + \cdots + \Delta L_n^2\right)^{1/2}/n \tag{13.57}$$

$$= \frac{\left(n\Delta L_0^2\right)^{1/2}}{n} \tag{13.58}$$

$$= \frac{\Delta L_0}{\sqrt{n}}, \tag{13.59}$$

where ΔL_0 is the statistical variation of the length for a transistor with width W_0. Equation (13.59) reveals that for a given W_0, as n increases, the variation of L_{eq} decreases.

The above result can be extended to other device parameters as well. For example, we postulate that μC_{ox} and V_{TH} suffer from less mismatch if the device area increases. Illustrated in Fig. 13.20, the reason is that a large transistor can be decomposed into a series and

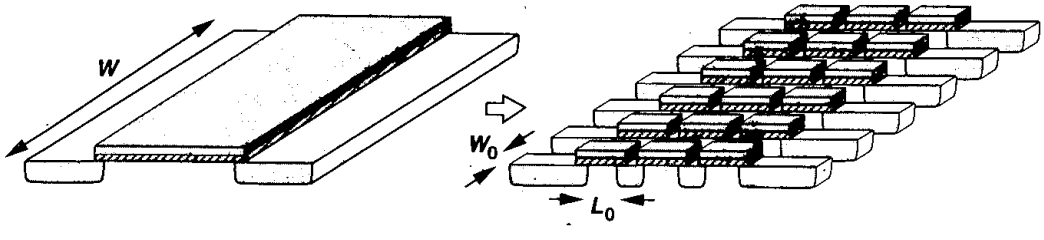


Figure 13.20 Large MOSFET viewed as a combination of small devices.

parallel combination of small unit transistors with dimensions W_0 and L_0, each exhibiting $(\mu C_{ox})_j$ and V_{THj}. For given W_0 and L_0, as the number of unit transistors increases, μC_{ox} and V_{TH} experience greater averaging, leading to smaller mismatch between two large transistors.

The foregoing qualitative observations have been verified mathematically and experimentally [2, 3]. Here, we state without proof that

$$\Delta V_{TH} = \frac{A_{VTH}}{\sqrt{WL}} \tag{13.60}$$

$$\Delta\left(\mu C_{OX}\frac{W}{L}\right) = \frac{A_K}{\sqrt{WL}}, \tag{13.61}$$

where A_{VTH} and A_K are proportionality factors.

Interestingly, A_{VTH} has been observed to scale down with the gate oxide thickness [3]. From the data in [4], $A_{VTH} \approx 10$ mV·μm for $t_{ox} \approx 100$ Å. Thus, in a 0.6-μm technology with $t_{ox} = 100$ Å, two 100 μm/0.6 μm devices ($L_{eff} \approx 0.5$ μm) exhibit a threshold mismatch of 1.4 mV. With this information, we can write

$$\Delta V_{TH} = \frac{0.1 t_{ox}}{\sqrt{WL}} \text{ mV}, \tag{13.62}$$

where t_{ox} is expressed in angstroms and W and L in microns. Since the channel capacitance is proportional to WLC_{ox}, we note that ΔV_{TH} and the channel capacitance bear a trade-off.

We now study the effect of device mismatch upon the performance of circuits. Mismatches lead to three significant phenomena: dc offsets, finite even-order distortion, and lower common-mode rejection. The last phenomenon was studied in Chapter 4.

DC Offsets Consider the differential pair shown in Fig. 13.21(a). With $V_{in} = 0$ and perfect symmetry, $V_{out} = 0$, but in the presence of mismatches, $V_{out} \neq 0$. We say the circuit suffers from a dc "offset" equal to the observed value of V_{out} when V_{in} is set to

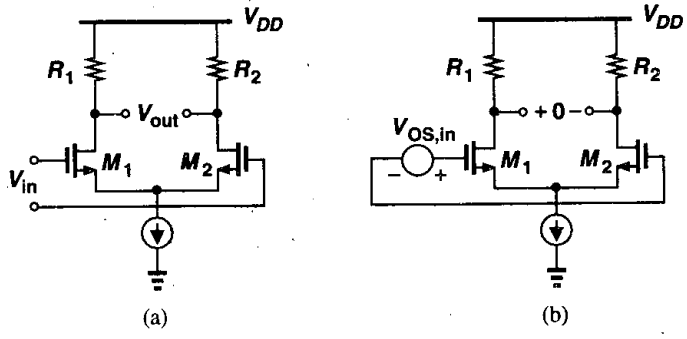


Figure 13.21 (a) Differential pair with offset measured at the output, (b) circuit of (a) with its offset referred to the input.

zero. In practice, it is more meaningful to specify the input-referred offset voltage, defined as the input level that forces the output voltage to go to zero [Fig. 13.21(b)]. Note that $|V_{OS,in}| = |V_{OS,out}|/A_v$. As with random noise, the polarity of random offsets is unimportant.

How does offset limit the performance? Suppose the differential pair of Fig. 13.21 is to amplify a small input voltage. Then, as depicted in Fig. 13.22, the output contains amplified replicas of both the signal and the offset. In a cascade of direct-coupled amplifiers, the dc offset may experience so much gain that it drives the latter stages into nonlinear operation.

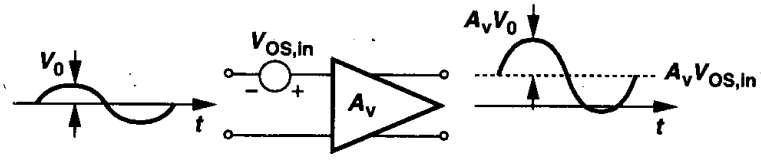


Figure 13.22 Effect of offset in an amplifier.

A more important effect of offset is the limitation on the precision with which signals can be measured. For example, if an amplifier is used to determine whether the input signal is greater or less than a reference, V_{REF} (Fig. 13.23), then the input-referred offset imposes a lower bound on the minimum $V_{in} - V_{REF}$ that can be detected reliably.

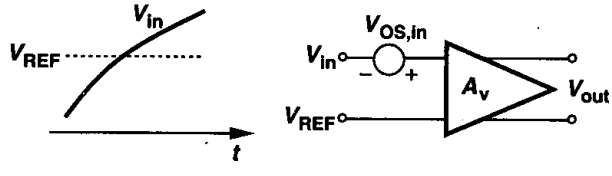


Figure 13.23 Accuracy limitation of an amplifier due to offset.

Let us now calculate the offset voltage of a differential pair, assuming that both the input transistors and the load resistors suffer from mismatch. As illustrated in Fig. 13.21(b), our objective is to find the value of $V_{OS,in}$ such that $V_{out} = 0$. The device mismatches are incorporated as $V_{TH1} = V_{TH}, V_{TH2} = V_{TH} + \Delta V_{TH}; (W/L)_1 = W/L, (W/L)_2 = W/L + \Delta(W/L); R_1 = R_D, R_2 = R_D + \Delta R$. For simplicity, $\lambda = \gamma = 0$, and mismatches in $\mu_n C_{ox}$ are neglected. For $V_{out} = 0$, we must have $I_{D1}R_1 = I_{D2}R_2$, concluding that I_{D1} cannot be equal to I_{D2}. Thus, we assume $I_{D1} = I_D, I_{D2} = I_D + \Delta I_D$.

Since $V_{OS,in} = V_{GS1} - V_{GS2}$, we have

$$V_{OS,in} = \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_1}} + V_{TH1} - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_2}} - V_{TH2} \tag{13.63}$$

$$= \sqrt{\frac{2}{\mu_n C_{ox}}}\left[\sqrt{\frac{I_D}{\frac{W}{L}}} - \sqrt{\frac{I_D + \Delta I_D}{\frac{W}{L} + \Delta\left(\frac{W}{L}\right)}}\right] - \Delta V_{TH} \tag{13.64}$$

$$= \sqrt{\frac{2}{\mu_n C_{ox}}}\sqrt{\frac{I_D}{W/L}}\left[1 - \sqrt{\frac{1 + \frac{\Delta I_D}{I_D}}{1 + \Delta\left(\frac{W}{L}\right)\Big/\left(\frac{W}{L}\right)}}\right] - \Delta V_{TH}. \tag{13.65}$$

Assuming $\Delta I_D/I_D$ and $\Delta(W/L)/(W/L) \ll 1$, and noting that for $\epsilon \ll 1$ we can write $\sqrt{1+\epsilon} \approx 1 + \epsilon/2$ and $(\sqrt{1+\epsilon})^{-1} \approx 1 - \epsilon/2$, we reduce (13.65) to

$$V_{OS,in} = \sqrt{\frac{2I_D}{\mu_n C_{ox}\left(\frac{W}{L}\right)}}\left\{1 - \left(1 + \frac{\Delta I_D}{2I_D}\right)\left[1 - \frac{\Delta(W/L)}{2(W/L)}\right]\right\} - \Delta V_{TH} \tag{13.66}$$

$$= \sqrt{\frac{2I_D}{\mu_n C_{ox}\left(\frac{W}{L}\right)}}\left[\frac{-\Delta I_D}{2I_D} + \frac{\Delta(W/L)}{2(W/L)}\right] - \Delta V_{TH}, \tag{13.67}$$

where the product of two small quantities is neglected. Recall that $I_{D1}R_1 = I_{D2}R_2$ and hence $I_D R_D = (I_D + \Delta I_D)(R_D + \Delta R_D) \approx I_D R_D + R_D \Delta I_D + I_D \Delta R_D$. Consequently, $\Delta I_D/I_D \approx -\Delta R_D/R_D$, and

$$V_{OS,in} = \frac{1}{2}\sqrt{\frac{2I_D}{\mu_n C_{ox}\left(\frac{W}{L}\right)}}\left[\frac{\Delta R_D}{R_D} + \frac{\Delta(W/L)}{(W/L)}\right] - \Delta V_{TH}. \tag{13.68}$$

We also recognize that the square-root quantity is approximately equal to the equilibrium overdrive voltage of each transistor, $V_{GS} - V_{TH}$, and

$$V_{OS,in} = \frac{V_{GS} - V_{TH}}{2}\left[\frac{\Delta R_D}{R_D} + \frac{\Delta(W/L)}{(W/L)}\right] - \Delta V_{TH}. \tag{13.69}$$

Equation (13.69) is an important result, revealing the dependence of $V_{OS,in}$ on device mismatches and bias conditions. We note that (1) the contribution of load resistor mismatch and transistor dimension mismatch *increases* with the equilibrium overdrive, and (2) the threshold voltage mismatch is directly referred to the input. Thus, it is desirable to minimize $V_{GS} - V_{TH}$ by lowering the tail current or increasing the transistor widths. In reality, since mismatches are independent statistical variables, we express (13.69) as[2]

$$V_{OS,in}^2 = \left(\frac{V_{GS} - V_{TH}}{2}\right)^2 \left\{\left(\frac{\Delta R_D}{R_D}\right)^2 + \left[\frac{\Delta(W/L)}{(W/L)}\right]^2\right\} + \Delta V_{TH}^2, \tag{13.70}$$

where squared quantities represent standard deviations.

To gain more insight into the effect of offset, let us establish an analogy between offset and *noise*. If the two inputs of a differential pair are shorted, the output voltage exhibits a finite noise, that is, a voltage that varies with time. We may therefore say that the offset voltage of a differential pair resembles a very low-frequency noise component, varying so slowly that it appears constant in our measurements. Viewed as such, offsets can be incorporated as noise sources, allowing us to utilize analysis techniques developed in Chapter 7. To this end, we represent the offset of two nominally-identical transistors by a voltage source equal to (13.70) in series with the gate of one of the transistors.

Example 13.3

Calculate the input-referred offset voltage of the circuit shown in Fig. 13.24(a). Assume all of the transistors operate in saturation.

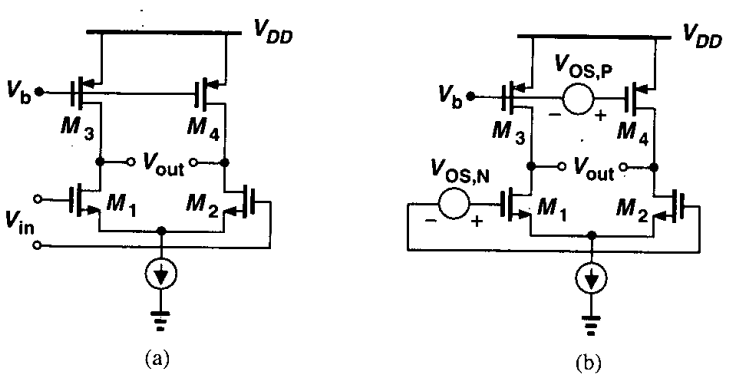


Figure 13.24

[2]As mentioned earlier, ΔV_{TH} does depend on W, an effect that can be added as a cross-correlation term. We neglect this term here for simplicity.

Solution

We insert the offsets of the NMOS and PMOS pairs as in Fig. 13.24(b). To obtain $I_{D1} = I_{D2}$ and $I_{D3} = I_{D4}$, we have from (13.69),

$$V_{OS,N} = \frac{(V_{GS} - V_{TH})_N}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_N + \Delta V_{TH,N} \tag{13.71}$$

$$V_{OS,P} = \frac{|V_{GS} - V_{TH}|_P}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_P + \Delta V_{TH,P}. \tag{13.72}$$

From the noise analysis in Chapter 7, $V_{OS,P}$ is amplified by a gain of $g_{mP}(r_{ON}||r_{OP})$ and divided by $g_{mN}(r_{ON}||r_{OP})$ when referred to the main input. As a result,

$$V_{OS,in} = \left\{\frac{|V_{GS} - V_{TH}|_P}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_P + \Delta V_{TH,P}\right\}\frac{g_{mP}}{g_{mN}} + \frac{(V_{GS} - V_{TH})_N}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_N + \Delta V_{TH,N}. \tag{13.73}$$

In practice, we add the "power" of these terms, as exemplified by (13.70). Note that, as with noise, the contribution of the offset of the PMOS pair is proportional to g_{mP}/g_{mN}.

The foregoing example can be better understood if we study the offset behavior of current sources. Consider the nominally-identical current sources M_1 and M_2 in Fig. 13.25. Neglecting channel-length modulation, we determine the total mismatch between I_{D1} and I_{D2} by calculating the total differential. Recall from calculus that if $y = f(x_1, x_2, \ldots)$, then the total differential is given by

$$\Delta y = \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \cdots. \tag{13.74}$$

Equation (13.74) simply means that each mismatch component Δx_j is weighted by the corresponding sensitivity $\partial f/\partial x_j$ as it contributes to the total mismatch. Since $I_D = (1/2)\mu_n C_{ox}(W/L)(V_{GS} - V_{TH})^2$, we have

$$\Delta I_D = \frac{\partial I_D}{\partial(W/L)}\Delta\left(\frac{W}{L}\right) + \frac{\partial I_D}{\partial(V_{GS} - V_{TH})}\Delta(V_{GS} - V_{TH}), \tag{13.75}$$

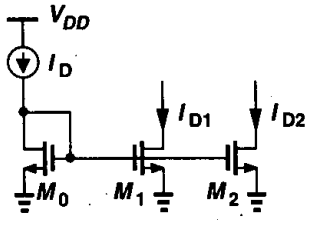


Figure 13.25 Mismatch between two current sources.

where mismatches in $\mu_n C_{ox}$ are neglected. It follows that

$$\Delta I_D = \frac{1}{2}\mu_n C_{ox}(V_{GS} - V_{TH})^2 \Delta\left(\frac{W}{L}\right) - \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})\Delta V_{TH}. \quad (13.76)$$

Unlike the input-referred offset *voltage*, current mismatch is usually normalized to the average value to allow a meaningful comparison:

$$\frac{\Delta I_D}{I_D} = \frac{\Delta(W/L)}{W/L} - 2\frac{\Delta V_{TH}}{V_{GS} - V_{TH}}. \quad (13.77)$$

This result suggests that, to minimize current mismatch, the overdrive voltage must be *maximized*, a trend opposite of that in (13.69). This is because as $V_{GS} - V_{TH}$ increases, threshold mismatch has lesser effect on the device currents.

The dependence of offset voltage and current mismatches upon the overdrive voltage is similar to our observations in Chapter 7 for corresponding noise quantities. For a given current, the input noise voltage of a differential pair increases as the overdrive increases because $g_m = 2I_D/(V_{GS} - V_{TH})$. Also, the output noise current of current sources is proportional to g_m and hence proportional to $V_{GS} - V_{TH}$.

Even-Order Distortion Our study of nonlinearity in Section 13.1 implies that, by virtue of odd symmetry, differential circuits are free from even-order distortion. In reality, however, mismatches degrade the symmetry, thereby introducing a finite even-order nonlinearity.

Analysis of the even-order distortion in the presence of mismatches is generally quite complex, often necessitating simulations. Here, we consider a simple case to gain some insight. Suppose the two signal paths in a differential circuit are represented by $y_1 \approx \alpha_1 x_1 + \alpha_2 x_1^2 + \alpha_3 x_1^3$ and $y_2 \approx \beta_1 x_2 + \beta_2 x_2^2 + \beta_3 x_2^3$ (Fig. 13.26). The differential output is given by

$$y_1 - y_2 = (\alpha_1 x_1 - \beta_2 x_2) + (\alpha_2 x_1^2 - \beta_2 x_2^2) + (\alpha_3 x_1^3 - \beta_3 x_2^3), \quad (13.78)$$

which, for $x_1 = -x_2$, reduces to

$$y_1 - y_2 = (\alpha_1 + \beta_1)x_1 + (\alpha_2 - \beta_2)x_1^2 + (\alpha_3 + \beta_3)x_1^3. \quad (13.79)$$

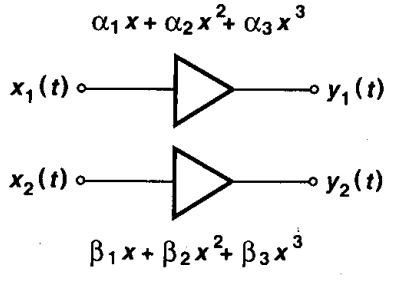


Figure 13.26 Effect of mismatch on second-order distortion.

If $x_1(t) = A\cos\omega t$, then the second harmonic has an amplitude equal to $(\alpha_2 - \beta_2)A^2/2$, i.e., proportional to the mismatch between the second-order coefficients of the input/output characteristic.

We should also mention that since at high frequencies, signals experience considerable phase shift, even-order distortion may arise from *phase* mismatch. This point is considered in Problem 13.1.

In circuits dissipating a high power, thermal gradients across the chip may create asymmetries. For example, if one transistor of a differential pair is closer to a high-power output stage than the other transistor, then mismatches arise between the threshold voltages and the mobilities of the two transistors.

13.2.1 Offset Cancellation Techniques

As mentioned above, the threshold voltage mismatch of MOSFETS trades with the channel capacitance. For example, a threshold mismatch of 1 mV translates to roughly 300 fF of channel capacitance for each transistor in a 0.6-μm technology. If many differential pairs are connected in parallel (e.g., in an A/D converter), the input capacitance becomes prohibitively large, severely degrading the speed and/or demanding high power dissipation in the preceding stage. Another difficulty is that mechanical stress may increase the offset voltages after a circuit is packaged. For these reasons, many high-precision systems require electronic cancellation of the offsets. As explained below, offset cancellation can also reduce $1/f$ noise of amplifiers considerably.

As our first step toward understanding the principle of offset cancellation, let us consider the circuit of Fig. 13.27(a), where a differential amplifier having an input-referred offset

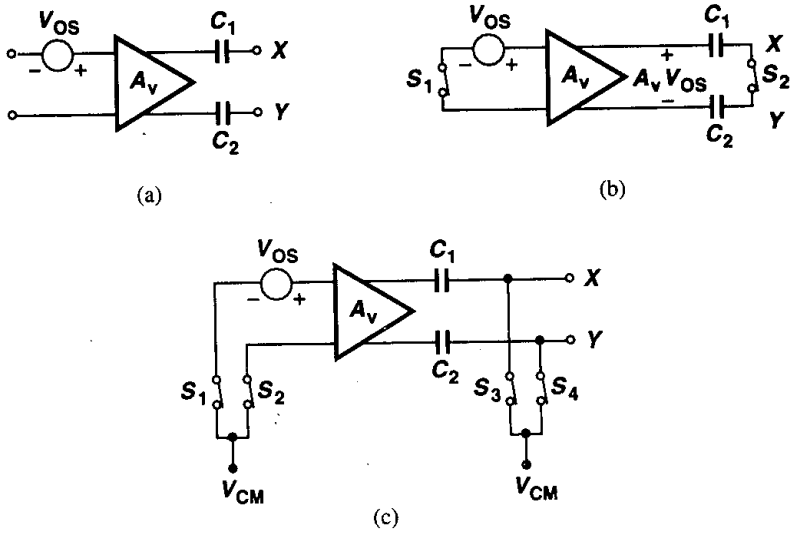


Figure 13.27 (a) Simple amplifier with capacitive coupling at the output, (b) circuit of (a) with its inputs and outputs shorted, (c) proper setting of the common-mode level during offset cancellation.

voltage V_{OS} is followed by two series capacitors. Now suppose, as shown in Fig. 13.27(b), the inputs are shorted together, driving the amplifier output to $V_{out} = A_v V_{OS}$. Furthermore, assume that during this period, nodes X and Y are shorted together as well. We note that when all of the node voltages are settled and $A_v V_{OS}$ is stored across C_1 and C_2, a zero differential input results in a zero difference between V_X and V_Y. Thus, after S_1 and S_2 turn off, the circuit consisting of the amplifier and C_1 and C_2 exhibits a zero offset voltage, amplifying only *changes* in the differential input voltage. In practice, the inputs and outputs must be shorted to proper common-mode voltages [Fig. 13.27(c)].

In summary, this type of offset cancellation "measures" the offset by setting the differential input to zero and stores the result on capacitors in series with the output. The circuit therefore requires a dedicated offset cancellation period, during which the actual input is disabled. Fig. 13.28 depicts the final topology, where CK denotes the offset cancellation command. Called "output offset storage," this technique reduces the overall offset to zero if S_3-S_4 exhibit no charge injection mismatch. Note, however, that if A_v is large, $A_v V_{OS}$ may "saturate" the amplifier output. For this reason, A_v is typically chosen to be less than roughly 10.

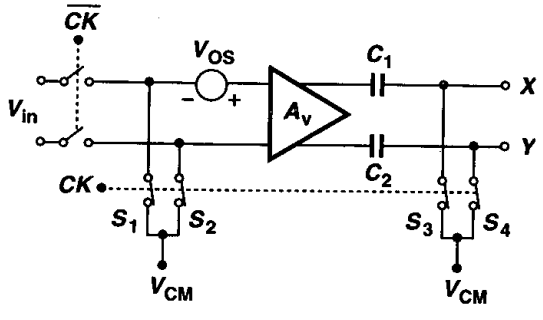


Figure 13.28 Control of amplification and offset cancellation modes by a clock.

In applications where a high voltage gain is required, the topology of Fig. 13.29(a) may be employed. Called "input offset storage," this approach incorporates two series capacitors at the input and places the amplifier in a unity-gain negative-feedback loop during offset cancellation. Thus, from Fig. 13.29(b), $V_{out} = V_{XY}$ and $(V_{out} - V_{OS})(-A_v) = V_{out}$. That is,

$$V_{out} = \frac{A_v}{1 + A_v} V_{OS} \tag{13.80}$$

$$\approx V_{OS}. \tag{13.81}$$

In essence, the circuit reproduces the amplifier's offset at nodes X and Y, storing the result on C_1 and C_2. Note that for a zero differential input, the differential output is equal to V_{OS}. Therefore, the input-referred offset voltage of the overall circuit (after S_3 and S_4 turn off) equals V_{OS}/A_v if S_3 and S_4 match perfectly (and the input capacitance of the amplifier is much less than C_1 and C_2). In reality, however, when S_3 and S_4 turn off, their charge injection mismatch may saturate the amplifier if A_v is very large.

The general drawback of input and output storage techniques is that they introduce capacitors in the signal path, a particularly serious issue in op amps and feedback systems.

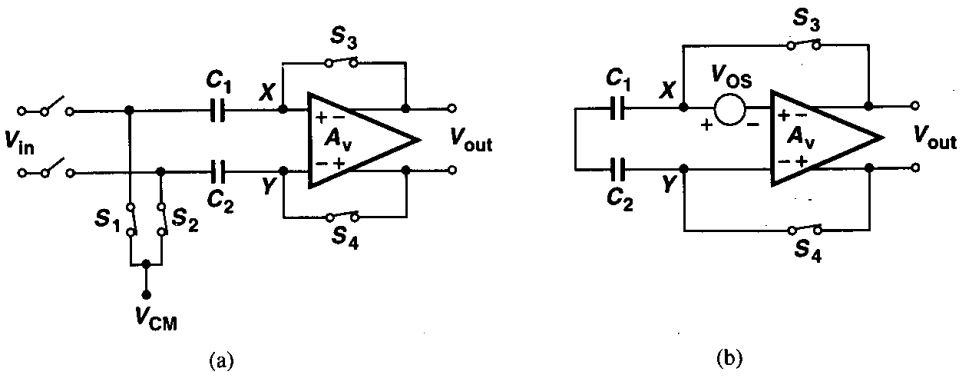


Figure 13.29 (a) Input offset storage, (b) circuit of (a) in the offset cancellation mode.

The bottom-plate parasitic of the capacitors may reduce the magnitude of the poles in the circuit, thereby degrading the phase margin. Even in open-loop amplifiers, this parasitic may limit the settling speed, intensifying the speed-power trade-off.

To resolve the above issues, the offset cancellation scheme can isolate the signal path from the offset storage capacitors though the use of an "auxiliary" amplifier. Consider the topology shown in Fig. 13.30, where A_{aux} amplifies the differential voltage V_1 stored across C_1 and C_2 and subtracts the result from the output of A_1. We note that if $V_{OS1}A_1 = V_1A_{aux}$, then for $V_{in} = 0$, $V_{out} = 0$, and the circuit is free from offsets. The key point here is that C_1 and C_2 do not appear in the signal path.

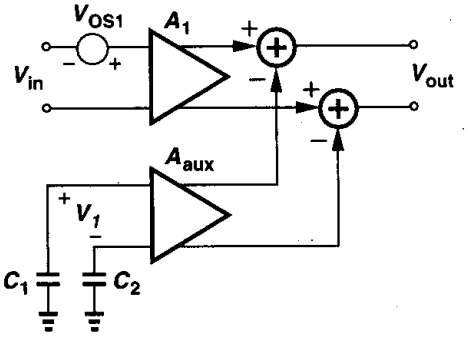


Figure 13.30 Addition of an auxiliary stage to remove the offset of an amplifier.

How is V_1 generated in Fig. 13.30? This is accomplished as illustrated in Fig. 13.31. Here, a second stage, A_2, is added and its output is sensed by A_{aux} during offset cancellation. To understand the operation, suppose that first only S_1 and S_2 are on, yielding $V_{out} = V_{OS1}A_1A_2$. Now, assume S_3 and S_4 turn on, placing A_2 and A_{aux} in a negative feedback loop. The reader can show that V_{out} then drops by a factor approximately equal to the loop gain: $V_{OS1}A_1A_2/(A_2A_{aux}) = V_{OS1}A_1/A_{aux}$. Stored across C_1 and C_2, this value is indeed the required V_1 in Fig. 13.30 because $(V_{OS1}A_1/A_{aux})A_{ux} = V_{OS1}A_1$.

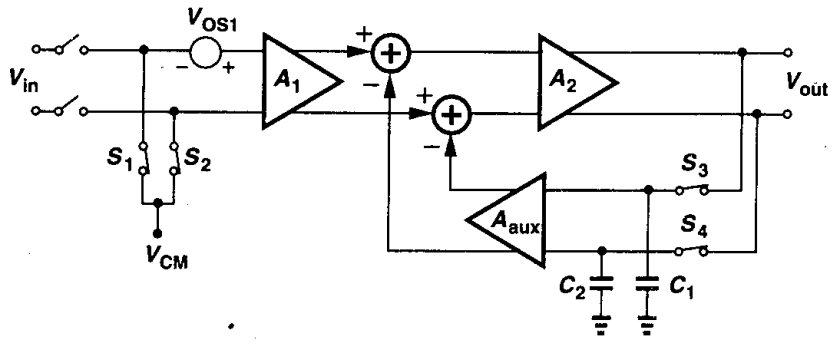


Figure 13.31 Auxiliary amplifier placed in a feedback loop during offset cancellation.

The topology of Fig. 13.31 suffers from two drawbacks. First, two voltage gain stages in the signal path may not be desirable in a high-speed op amp. Second, addition of the output voltages of A_1 and A_2 is quite difficult. For these reasons, the technique is usually realized as shown in Fig. 13.32(a), where each G_m stage is simply a differential pair and the R stage represents a transimpedance amplifier. As exemplified by Fig. 13.32(b), G_{m1} and R may in fact constitute a one-stage op amp while G_{m2} adds an offset correction current at the low impedance nodes X and Y.

Let us now examine the offset cancellation in Fig. 13.32(a) carefully, taking the offset voltage of G_{m2} into account as well. As depicted in Fig. 13.33, we can write:

$$[G_{m1}V_{OS1} - G_{m2}(V_{out} - V_{OS2})]R = V_{out}. \tag{13.82}$$

Thus,

$$V_{out} = \frac{G_{m1}RV_{OS1} + G_{m2}RV_{OS2}}{1 + G_{m2}R}. \tag{13.83}$$

This voltage is stored on C_1 and C_2 after S_3 and S_4 turn off. The offset voltage referred to the main input is therefore given by

$$V_{OS,tot} = \frac{V_{out}}{G_{m1}R} \tag{13.84}$$

$$= \frac{V_{OS1}}{1 + G_{m2}R} + \frac{G_{m2}}{G_{m1}}\frac{V_{OS2}}{1 + G_{m2}R} \tag{13.85}$$

$$\approx \frac{V_{OS1}}{G_{m2}R} + \frac{V_{OS2}}{G_{m1}R}, \tag{13.86}$$

where we have assumed $G_{m2}R \gg 1$. If $G_{m2}R$ and $G_{m1}R$ are large, as in the op amp of Fig. 13.32(b), then $V_{OS,tot}$ is very small.

The offset cancellation of Fig. 13.32 warrants a cautionary note. Upon turning off, S_3 and S_4 may inject slightly unequal charges onto C_1 and C_2, respectively, creating an error

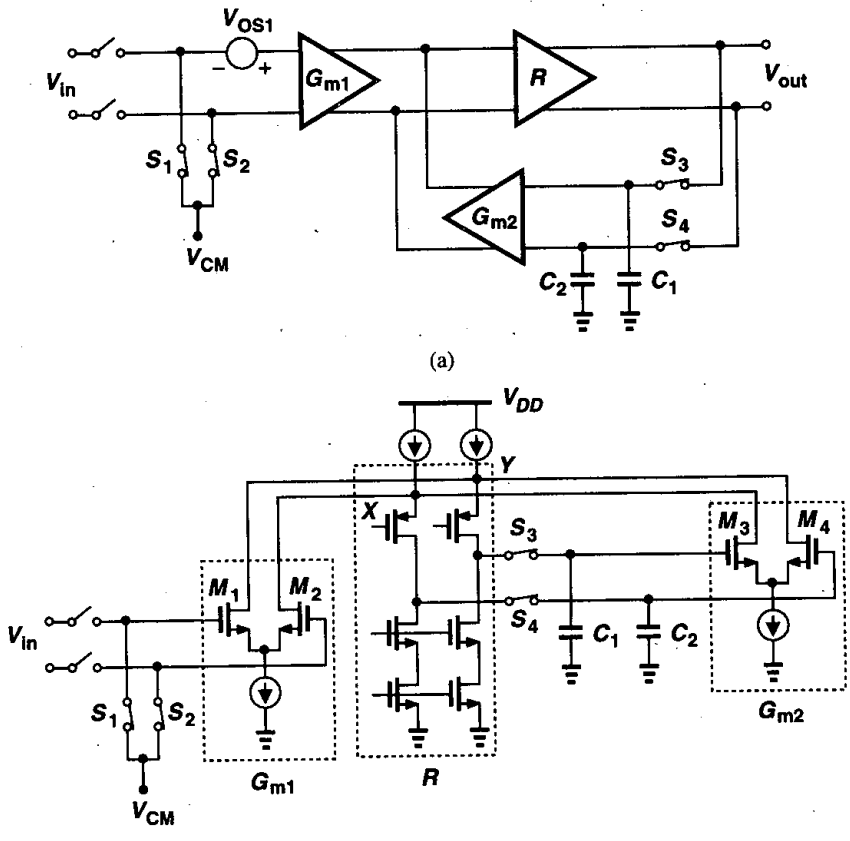


Figure 13.32 (a) Circuit of Fig. 13.31 using G_m and R stages, (b) realization of (a) in a folded-cascode op amp.

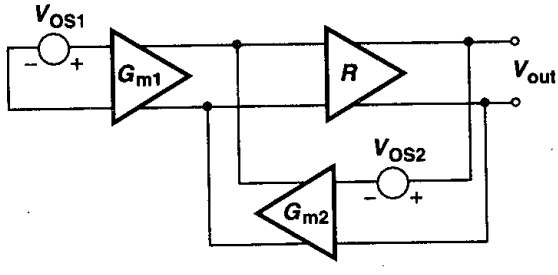


Figure 13.33 Circuit of Fig. 13.32(a) including offset of G_{m2}.

voltage that is *not* corrected because the feedback loop is opened. The reader can prove thatfor a differential injection-induced error voltage of ΔV, the resulting input-referred offset voltage equals $(G_{m2}/G_{m1})\Delta V$. For this reason, G_{m2} is usually chosen to be on the order of $0.1G_{m1}$.

We should also mention that the unity-gain and precision multiply-by-two circuits described in Chapter 12 cancel the offset of the op amp as well. The proof is left to the reader.[3]

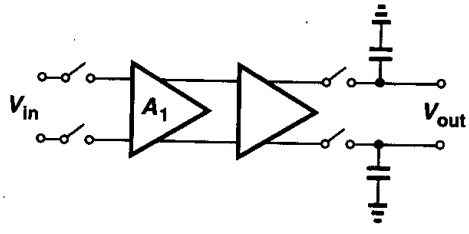
It is important to note that the offset cancellation techniques studied here require periodic refreshing because the junction and subthreshold leakage of the switches eventually corrupts the correction voltage stored across the capacitors. In a typical design, the offset must be refreshed at a rate of at least a few kilohertz.

13.2.2 Reduction of Noise by Offset Cancellation

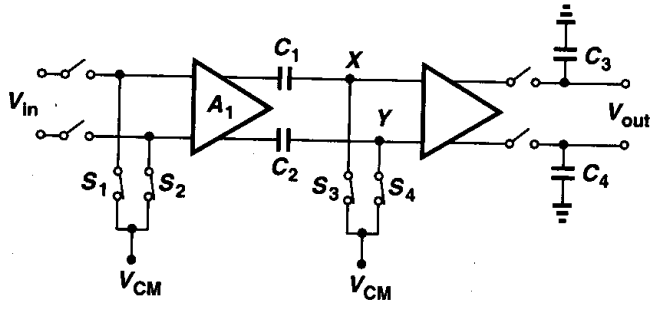
Recall from previous sections that the offset of a differential amplifier can be viewed as a noise component having a very low frequency. We therefore expect that periodic offset cancellation can potentially reduce the (low-frequency) noise of the circuit as well.

Consider a simple differential amplifier that is to be used in the front-end of a sam-

[3]If, as shown in Fig. 12.34, an equalizing switch is added to the circuit, then the op amp offset may not be removed.



(a)



(b)

Figure 13.34 (a) Front end of a sampler, (b) circuit of (a) with offset cancellation applied to the first stage.

pling system [Fig. 13.34(a)]. Here, the noise of A_1 directly corrupts V_{in}. The $1/f$ noise of A_1 proves especially problematic if the signal spectrum extends from zero to only a few megahertz, because the $1/f$ noise corner frequency is typically around 500 kHz to 1 MHz.

Now suppose the amplifier undergoes offset cancellation before *every* sampling operation [Fig. 13.34(b)]. That is, as depicted in Fig. 13.35, the input is disabled; the offset of A_1 is stored on C_1 and C_2; the input is enabled and amplified by A_1 and A_2 and stored on C_3 and C_4; and finally the sampling switches are turned off. How does the noise of A_1 affect the final output? Denoting the time elapsed from the end of offset cancellation to the end of sampling by $\Delta t = t_2 - t_1$, we recall that at $t = t_1$, $V_{XY} = 0$. Thus, from t_1 to t_2, only *high-frequency* noise components of A_1, on the order of $> 1/\Delta t$, change V_{XY} significantly. In other words, offset cancellation suppresses noise frequencies below roughly $1/\Delta t$.

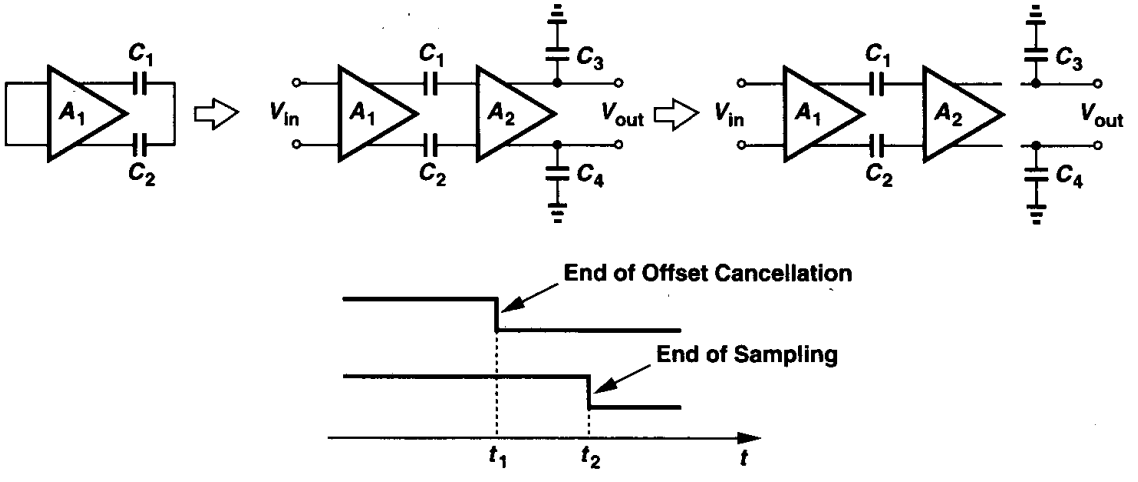


Figure 13.35 Sequence of operations in the sampler.

To better understand this concept, let us consider a numerical example. Assuming $\Delta t = 10$ ns, we examine two noise components, one at 1 MHz and another at 10 MHz, approximating each with a sinusoid (Fig. 13.36). For a sinusoid of amplitude A and frequency f, the maximum slew rate is equal to $2\pi f A$ and hence the maximum variation in Δt seconds is $2\pi f A \Delta t$. Normalizing this value to the amplitude, we obtain the change for 1-MHz and 10-MHz components as $\Delta V_1/A = 6.3\%$ and $\Delta V_2/A = 63\%$, respectively. We therefore conclude that noise frequencies below a few megahertz do not have sufficient time to change if the sampling occurs only 10 ns after the end of offset cancellation.

Originally utilized in charge-coupled devices (CCDs), the foregoing property of offset cancellation is called "correlated double sampling" (CDS) because it involves two consecutive sampling operations (the first being offset storage) that are so tightly spaced in time that they do not allow (low-frequency) noise components to vary significantly. A powerful technique, CDS finds wide usage in suppressing the $1/f$ noise of MOS circuits. Nonetheless, it leads to aliasing of wideband noise [5].

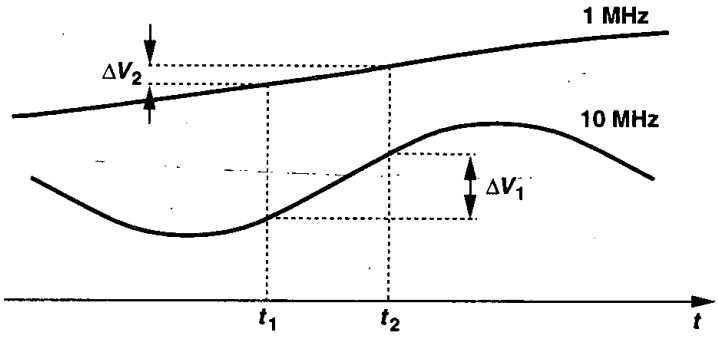


Figure 13.36 Variation of 1-MHz and 10-MHz noise components in a time interval of 10 ns.

13.2.3 Alternative Definition of CMRR

Recall from Chapter 4 that common-mode rejection is represented by the change in the differential output divided by the change in the input common-mode level and CMRR is defined as the differential gain divided by this quantity. We also noted that in fully differential circuits, the finite output impedance of the tail current source and asymmetries limit the common-mode rejection.

Now consider a differential circuit sensing an input CM change, $\Delta V_{in,CM}$. If the differential output voltage changes by ΔV_{out} while the differential input voltage is zero, we can say that the output *offset* voltage of the circuit has changed by ΔV_{out}. In other words, common-mode rejection can be viewed as the change in the output offset divided by the change in the input CM level. Following the notation in Chapter 4, we write

$$A_{CM-DM} = \frac{\Delta V_{OS,out}}{\Delta V_{CM,in}}. \tag{13.87}$$

Since $CMRR = A_{DM}/A_{CM-DM}$, we have

$$CMRR = \frac{A_{DM}}{\dfrac{\Delta V_{OS,out}}{\Delta V_{CM,in}}} \tag{13.88}$$

$$= \frac{\dfrac{\Delta V_{CM,in}}{\Delta V_{OS,out}}}{A_{DM}}. \tag{13.89}$$

Noting that $\Delta V_{OS,out}/A_{DM}$ is in fact the input-referred offset voltage, we have

$$CMRR = \frac{\Delta V_{CM,in}}{\Delta V_{OS,in}}. \tag{13.90}$$

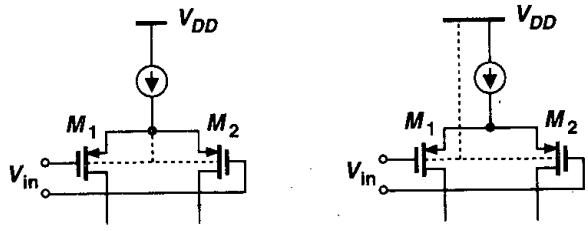


Figure 13.37 PMOS differential pair (a) without and (b) with body effect.

The above result proves useful in analyzing the behavior of circuits. For example, suppose an op amp incorporates a PMOS differential pair at the input. Which one of the topologies shown in Fig. 13.37 yields a higher CMRR? In Fig. 13.37(a), body effect is eliminated and the threshold voltages of M_1 and M_2 are independent of the input CM level. In Fig. 13.37(b), on the other hand, M_1 and M_2 experience body effect and, if they suffer from mismatches in their body effect coefficients, then the difference between V_{TH1} and V_{TH2}, i.e., the input offset voltage, *varies* with the input CM level, degrading the common-mode rejection.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume $V_{DD} = 3$ V where necessary. Also, assume all transistors are in saturation.

13.1. The input-output characteristic of an amplifier is approximated as $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t)$ in the range $x = [0 \ x_{max}]$.
(a) What is the maximum nonlinearity?
(b) What is the THD for $x(t) = (x_{max} \cos \omega t + x_{max})/2$.

13.2. In the circuits of Fig. 13.6, $W/L = 20/0.5$ and $I = 0.5$ mA. Calculate the harmonic distortion in each circuit if the input signal has a peak amplitude of 100 mV. How do the results change if we double W/L or I?

13.3. For the circuits of Fig. 13.6(a), plot the THD and the input-referred thermal noise as a function of **(a)** W/L, **(b)** I. Identify the trade-offs between noise, linearity, and power dissipation.

13.4. In Fig. 13.6, *two* effects lead to a trade-off between nonlinearity and voltage gain. Describe these effects.

13.5. The circuit of Fig. 13.6(a) is designed with $W/L = 50/0.5$, $I = 1$ mA, and $R_D = 2$ kΩ. The circuit is placed in a feedback loop similar to that of Fig. 13.7 with $\beta = 0.2$ and senses an input sinusoid with a peak amplitude of 10 mV. Calculate the THD at the output.

13.6. Suppose in Fig. 13.14, A_1 and A_2 have an input-referred noise voltage V_n. Neglecting other sources of noise, calculate the input-referred noise voltage of the overall circuit.

13.7. Equation 13.36 suggests that if the open-loop gain, α_1, increases while other parameters remain constant, then the harmonic distortion drops sharply. Repeat Problem 13.5 with $W/L = 200/0.5$ to achieve a higher open-loop gain and explain the results.

13.8. Equation 13.36 suggests that if $\beta\alpha_1 \gg 1$, then $b/a \propto \beta^{-2}$. Repeat Problem 13.5 with $\beta = 0.4$.

13.9. Suppose the nonlinear feedforward amplifier in Fig. 13.7 is characterized by $y(t) = \alpha_1 x(t) + \alpha_3 x^3(t)$. Estimate the magnitude of the third harmonic at the output of the overall system.

13.10. As mentioned in Chapter 2, MOS devices operating in the subthreshold region exhibit an exponential behavior: $I_D = I_0 \exp[V_{GS}/(\zeta V_T)]$. Suppose both of the circuits shown in Fig. 13.6 operate in the subthreshold region. Derive expressions for the harmonic amplitudes if the input signal is much less than ζV_T. For the differential pair, first prove that $I_{D1} - I_{D2} \propto \tanh[V_{in}/(2\zeta V_T)]$ and then write the Taylor expansion of the hyperbolic tangent.

13.11. The mobility of MOSFETs is in fact a function of the gate-source voltage and expressed as $\mu = \mu_0/[1 + \theta(V_{GS} - V_{TH})]$, where θ is an empirical factor (Chapter 16). Assuming $\theta(V_{GS} - V_{TH}) \ll 1$ and using the relationship $(1+\epsilon)^{-1} \approx 1 - \epsilon$ for $\epsilon \ll 1$, calculate the third harmonic in the circuit of Fig. 13.6(a).

13.12. The input devices of a differential pair have an effective length of 0.5 μm.

(a) Assuming $\Delta V_{TH} = 0.1 t_{ox}/\sqrt{WL}$ and neglecting other mismatches, determine the minimum width of the transistors such that $V_{OS} \leq 5$ mV.

(b) If the tail current is 1 mA, what is the maximum input swing that gives a THD of 1%?

13.13. Repeat Problem 13.12(b) if the tolerable input offset is 2 mV and compare the results.

13.14. Determine the dimensions of M_1 and M_2 in Fig. 13.25 such that $I_{D1} \approx I_{D2} = 0.5$ mA, $\Delta I_D/I_D = 2\%$, and $V_{GS} - V_{TH} = 0.5$ V. Assume $\Delta V_{TH} = 0.1 t_{ox}/\sqrt{WL}$ and neglect other mismatches.

13.15. Source degeneration can improve the matching between current sources if resistor mismatches are small. Prove that in the circuit of Fig. 13.38,

$$\frac{\Delta I_D}{I_D} = \frac{1}{1 + g_m R_S}\left[\frac{\Delta(\mu_n C_{ox})}{\mu_n C_{ox}} + \frac{\Delta(W/L)}{(W/L)} - \frac{2\Delta V_{TH}}{V_{GS} - V_{TH}} - g_m \Delta R_S\right], \quad (13.91)$$

where ΔR_S denotes the mismatch between R_{S1} and R_{S2}. Note that for an appreciable reduction of $\Delta I/I_D$, R_S must be greater than $1/g_m$.

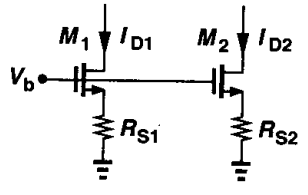


Figure 13.38

13.16. In the circuit of Fig. 13.26, assume $\alpha_j = \beta_j$ but $x_1(t) = A\cos\omega t$ and $x_2(t) = A\cos(\omega t + \theta)$, where θ denotes a small phase mismatch. Calculate the magnitude of the second harmonic at the output.

13.17. In the circuit of Fig. 13.39, M_3 and M_4 suffer from a threshold mismatch of ΔV_{TH} and the circuit is otherwise symmetric. Assuming $\lambda \neq 0$ but $\gamma = 0$, calculate the input-referred offset voltage. What happens as $R_D \to \infty$?

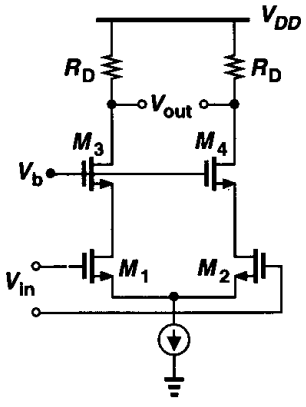


Figure 13.39

13.18. In the circuit of Fig. 13.29, the amplifier has an input capacitance (between X and Y) equal to C_{in}. Calculate the input offset voltage after offset compensation.

13.19. The circuit of Fig. 13.29 is designed for an input offset voltage of 1 mV. If the width of the transistors in the input differential pair of the amplifier is doubled, what is the overall input offset voltage? (Neglect the input capacitance of the amplifier.)

13.20. Explain why the circuit of Fig. 13.24 suffers from a trade-off between the input offset and the output voltage swing (for a given tail current).

References

1. F. Krummenacher and N. Joehl, "A 4-MHz CMOS Continuous-Time Filter with On-Chip Automatic Tuning," *IEEE J. Solid-State Circuits*, vol. 23, pp. 750–758, June 1988.
2. K. R. Lakshmikumar, R. A. Hadaway, and M. A. Copeland, "Characterization and Modeling of Mismatches in MOS Transistors for Precision Analog Design," *IEEE J. Solid-State Circuits*, vol. 21, pp. 1057–1066, Dec. 1986.
3. M. J. M. Pelgrom, A. C. J. Duinmaiger, and A. P. G. Welbers, "Matching Properties of MOS Transistors," *IEEE J. Solid-State Circuits*, vol. SC-24, pp. 1433–1439, Oct. 1989.
4. M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor Matching in Analog CMOS Applications," *IEDM Dig. of Tech. Papers*, pp. 34.1.1–34.1.4, Dec. 1998.
5. C. C. Enz and G. C. Temes, "Circuit Techniques for Reducing the Effects of Op-Amp Imperfections: Autozeroing, Correlated Double Sampling, and Chopper Stabilization," *Proc. IEEE*, vol. 84, pp. 1584–1614, Nov. 1996.

Chapter 14

Oscillators

Oscillators are an integral part of many electronic systems. Applications range from clock generation in microprocessors to carrier synthesis in cellular telephones, requiring vastly different oscillator topologies and performance parameters. Robust, high-performance oscillator design in CMOS technology continues to pose interesting challenges. As described in Chapter 15, oscillators are usually embedded in a phase-locked system.

This chapter deals with the analysis and design of CMOS oscillators, more specifically, voltage-controlled oscillators (VCOs). Beginning with a general study of oscillation in feedback systems, we introduce ring oscillators and LC oscillators along with methods of varying the frequency of oscillation. We then describe a mathematical model of VCOs that will be used in the analysis of PLLs in Chapter 15.

14.1 General Considerations

A simple oscillator produces a periodic output, usually in the form of voltage. As such, the circuit has no input while sustaining the output indefinitely. How can a circuit oscillate? Recall from Chapter 10 that negative feedback systems may oscillate, i.e., an oscillator is a badly-designed feedback amplifier![1] Consider the unity-gain negative feedback circuit shown in Fig. 14.1, where

$$\frac{V_{out}}{V_{in}}(s) = \frac{H(s)}{1 + H(s)}. \tag{14.1}$$

As mentioned in Chapter 10, if the amplifier itself experiences so much phase shift at high frequencies that the overall feedback becomes positive, then oscillation may occur. More accurately, if for $s = j\omega_0$, $H(j\omega_0) = -1$, then the closed-loop gain approaches infinity at ω_0. Under this condition, the circuit amplifies its own noise components at ω_0 indefinitely. In fact, as conceptually illustrated in Fig. 14.2, a noise component at ω_0 experiences a total gain of unity and a phase shift of 180°, returning to the subtractor as a negative replica

[1] It is said, "In the high-frequency world, amplifiers oscillate and oscillators don't."

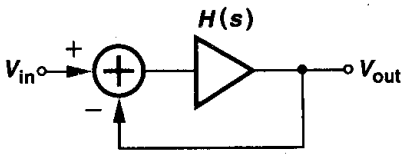


Figure 14.1 Feedback system.

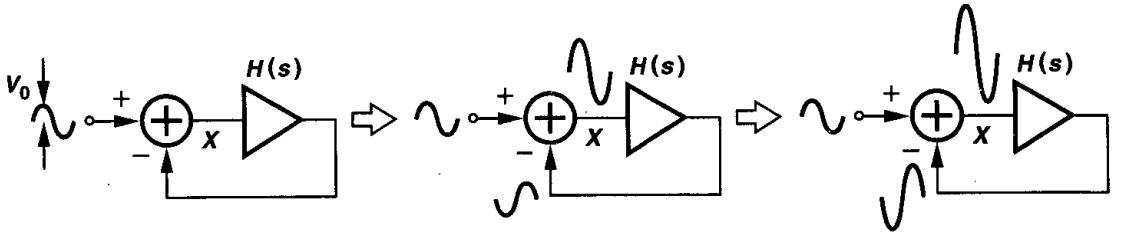


Figure 14.2 Evolution of oscillatory system with time.

of the input. Upon subtraction, the input and the feedback signals give a larger difference. Thus, the circuit continues to "regenerate," allowing the component at ω_0 to grow.

For the oscillation to begin, a loop gain of unity or greater is necessary. This can be seen by following the signal around the loop over many cycles and expressing the amplitude of the subtractor's output in Fig. 14.2 as a geometric series (if $\angle H(j\omega_0) = 180°$):

$$V_X = V_0 + |H(j\omega_0)|V_0 + |H(j\omega_0)|^2 V_0 + |H(j\omega_0)|^3 V_0 + \cdots. \tag{14.2}$$

If $|H(j\omega_0)| > 1$, the above summation diverges whereas if $|H(j\omega_0)| < 1$, then

$$V_X = \frac{V_0}{1 - |H(j\omega_0)|} < \infty. \tag{14.3}$$

In summary, if a negative-feedback circuit has a loop gain that satisfies two conditions:

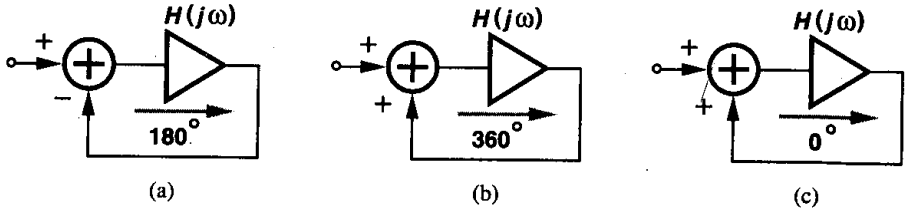
$$|H(j\omega_0)| \geq 1 \tag{14.4}$$

$$\angle H(j\omega_0) = 180°, \tag{14.5}$$

then the circuit may oscillate at ω_0. Called "Barkhausen criteria," these conditions are necessary but not sufficient [1]. In order to ensure oscillation in the presence of temperature and process variations, we typically choose the loop gain to be at least twice or three times the required value.

We may state the second Barkhausen criterion as $\angle H(j\omega) = 180°$ or a *total* phase shift of 360°. This should not be confusing: if the system is designed to have a low-frequency negative feedback, it already produces 180° of phase shift in the signal traveling around the loop (as represented by the subtractor in Fig. 14.1), and $\angle H(j\omega) = 180°$ denotes an additional *frequency-dependent* phase shift that, as illustrated in Fig. 14.2, ensures the

feedback signal *enhances* the original signal. Thus, the three cases illustrated in Fig. 14.3 are equivalent in terms of the second criterion. We say the system of Fig. 14.3(a) exhibits



(a) (b) (c)

Figure 14.3 Various views of oscillatory feedback system.

a frequency-dependent phase shift of 180° (denoted by the arrow) and a dc phase shift of 180°. The difference between Figs. 14.3(b) and (c) is that the open-loop amplifier in the former contains enough stages with proper polarities to provide a total phase shift of 360° at ω_0 whereas that in the latter produces *no* phase shift at ω_0. Examples of these topologies are presented later in this chapter.

CMOS oscillators in today's technology are typically implemented as "ring oscillators" or "LC oscillators." We study each type in the following sections.

14.2 Ring Oscillators

A ring oscillator consists of a number of gain stages in a loop. To arrive at the actual implementation, we begin by attempting to make a single-stage feedback circuit oscillate.

Example 14.1

Explain why a single common-source stage does not oscillate if it is placed in a unity-gain loop.

Solution

From Fig. 14.4, it is seen that the open-loop circuit contains only one pole, thereby providing a maximum frequency-dependent phase shift of 90° (at a frequency of infinity). Since the common-source stage exhibits a dc phase shift of 180° due to the signal inversion from the gate to the drain, the maximum total phase shift is 270°. The loop therefore fails to sustain oscillation growth.

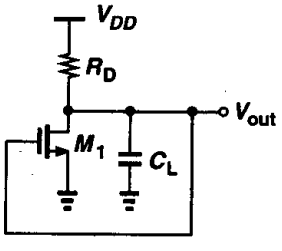


Figure 14.4

The above example suggests that oscillation may occur if the circuit contains multiple stages and hence multiple poles. Indeed, such a topology was considered *undesirable* in Chapter 10 because it led to inadequate phase margin in op amps. We therefore surmise that if the circuit of Fig. 14.4 is modified as shown in Fig. 14.5, then two significant poles appear in the signal path, allowing the frequency-dependent phase shift to approach 180°.

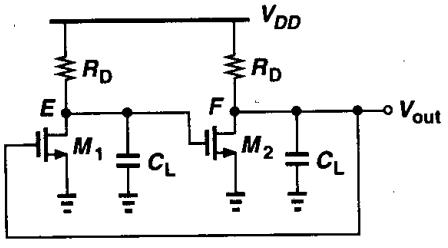


Figure 14.5 Two-pole feedback system.

Unfortunately, this circuit exhibits *positive* feedback near zero frequency due to the signal inversion through each common-source stage. As a result, it simply "latches up" rather than oscillates. That is, if V_E rises, V_F falls, thereby turning M_1 off and allowing V_E to rise further. This may continue until V_E reaches V_{DD} and V_F drops to near zero, a state that will remain indefinitely.

To gain more insight into the oscillation conditions, let us assume an ideal inverting stage (with zero phase shift at all frequencies) is inserted in the loop of Fig. 14.5, providing *negative* feedback near zero frequency and eliminating the problem of latch-up (Fig. 14.6). Does this circuit oscillate? We note that the loop contains only two poles: one at E and another

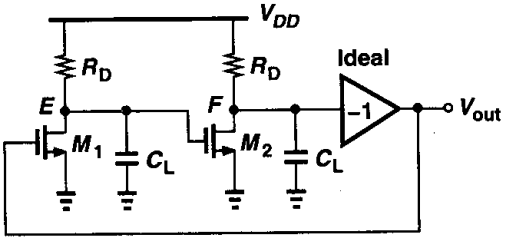


Figure 14.6 Two-pole feedback system with additional signal inversion.

at F. The frequency-dependent phase shift can therefore reach 180°, but at a frequency of infinity. Since the loop gain vanishes at very high frequencies, we observe that the circuit does not satisfy both of Barkhausen's criteria at the same frequency (Fig. 14.7), failing to oscillate.

The foregoing discussion points to the need for greater phase shift around the loop, suggesting the possibility of oscillation if the third inverting stage in Fig. 14.6 contains a pole that contributes significant phase. We then arrive at the topology depicted in Fig. 14.8. If the three stages are identical, the total phase shift around the loop, ϕ, reaches $-135°$ at $\omega = \omega_{p,E} (= \omega_{p,F} = \omega_{p,G})$ and $-270°$ at $\omega = \infty$. Consequently, ϕ equals $-180°$ at $\omega < \infty$, where the loop gain can be still greater than or equal to unity. This circuit indeed oscillates if the loop gain is sufficient and it is an example of a ring oscillator.

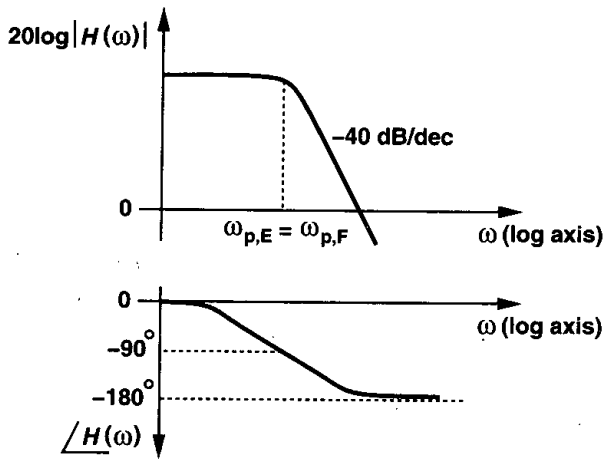


Figure 14.7 Loop gain characteristics of a two-pole system.

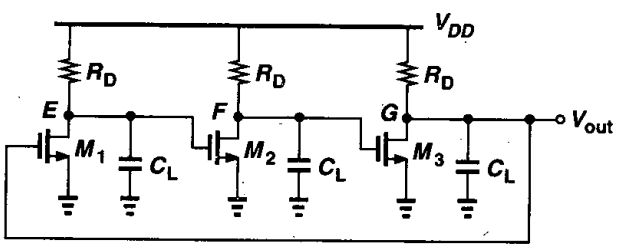


Figure 14.8 Three-stage ring oscillator.

It is instructive to calculate the minimum voltage gain per stage in Fig. 14.8 that is necessary for oscillation. Neglecting the effect of the gate-drain overlap capacitance and denoting the transfer function of each stage by $-A_0/(1+s/\omega_0)$, we have for the loop gain:

$$H(s) = -\frac{A_0^3}{(1+\frac{s}{\omega_0})^3}. \tag{14.6}$$

The circuit oscillates only if the frequency-dependent phase shift equals 180°, i.e., if each stage contributes 60°. The frequency at which this occurs is given by

$$\tan^{-1}\frac{\omega_{osc}}{\omega_0} = 60° \tag{14.7}$$

and hence:

$$\omega_{osc} = \sqrt{3}\omega_0. \tag{14.8}$$

The minimum voltage gain per stage must be such that the magnitude of the loop gain at ω_{osc} is equal to unity:

$$\frac{A_0^3}{\left[\sqrt{1+(\frac{\omega_{osc}}{\omega_0})^2}\right]^3} = 1. \tag{14.9}$$

It follows from (14.8) and (14.9) that

$$A_0 = 2. \tag{14.10}$$

In summary, a three-stage ring oscillator requires a low-frequency gain of 2 per stage, and it oscillates at a frequency of $\sqrt{3}\omega_0$, where ω_0 is the 3-dB bandwidth of each stage.

Let us now examine the waveforms at the three nodes of the oscillator of Fig. 14.8. Since each stage contributes a frequency-dependent phase shift of 60° as well as a low-frequency signal inversion, the waveform at each node is 240° (or 120°) out of phase with respect to its neighboring nodes (Fig. 14.9). The ability to generate multiple phases is a very useful property of ring oscillators.

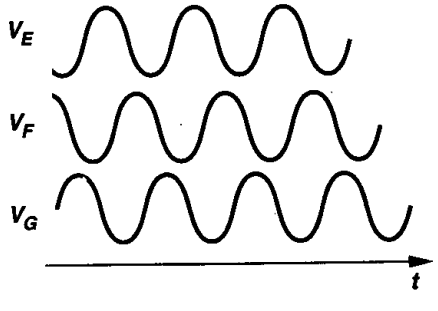


Figure 14.9 Waveforms of a three-stage ring oscillator.

Amplitude Limiting The natural question at this point is: what happens if in the three-stage ring of Fig. 14.8, $A_0 \neq 2$? We know from Barkhausen's criteria that if $A_0 < 2$, the circuit fails to oscillate, but what if $A_0 > 2$? To answer this question, we first model the oscillator by a linear feedback system, as depicted in Fig. 14.10. Note that the feedback is positive (i.e., V_{out} is *added* to V_{in}) because $H(s)$ in Eq. (14.6) already includes the negative polarity resulting from three inversions in the signal path. The closed-loop transfer function is:

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{\frac{-A_0^3}{(1+s/\omega_0)^3}}{1+\frac{A_0^3}{(1+s/\omega_0)^3}} \tag{14.11}$$

$$= \frac{-A_0^3}{(1+s/\omega_0)^3 + A_0^3}. \tag{14.12}$$

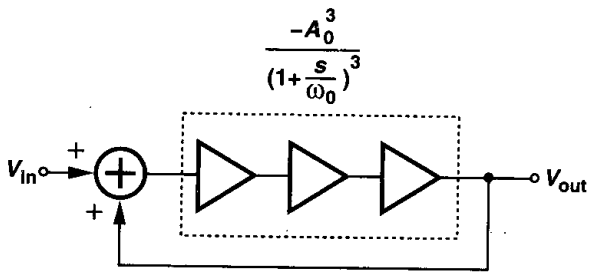


Figure 14.10 Linear model of three-stage ring oscillator.

The denominator of (14.12) can be expanded as:

$$(1+\frac{s}{\omega_0})^3 + A_0^3 = (1+\frac{s}{\omega_0}+A_0)\left[(1+\frac{s}{\omega_0})^2 - (1+\frac{s}{\omega_0})A_0 + A_0^2\right]. \tag{14.13}$$

Thus, the closed-loop system exhibits three poles:

$$s_1 = (-A_0 - 1)\omega_0 \tag{14.14}$$

$$s_{2,3} = [\frac{A_0(1 \pm j\sqrt{3})}{2} - 1]\omega_0. \tag{14.15}$$

Since A_0 itself is positive, the first pole leads to a decaying exponential term: $\exp[(-A_0 - 1)\omega_0 t]$, which can be neglected in the steady state. Figure 14.11 illustrates the locations of the poles for different values of A_0, revealing that for $A_0 > 2$, the two complex poles exhibit a positive real part and hence give rise to a growing sinusoid. Neglecting the effect of s_1, we express the output waveform as

$$V_{out}(t) = a\exp(\frac{A_0 - 2}{2}\omega_0 t)\cos(\frac{A_0\sqrt{3}}{2}\omega_0 t). \tag{14.16}$$

Thus, if $A_0 > 2$, the exponential envelope grows to infinity.

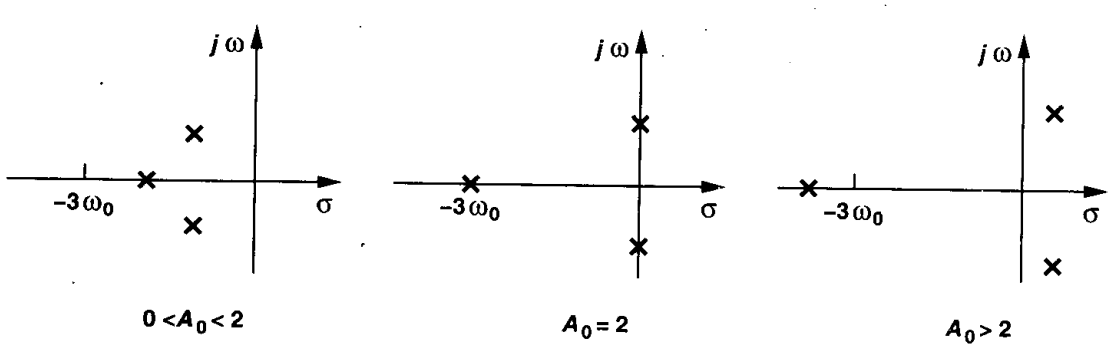


Figure 14.11 Poles of three-stage ring oscillator for various values of gain.

In practice, as the oscillation amplitude increases, the stages in the signal path experience nonlinearity and eventually "saturation," limiting the maximum amplitude. We may say the poles begin in the right half plane and eventually move to the imaginary axis to stop the growth. If the small-signal loop gain is greater than unity, the circuit must spend enough time in saturation so that the "average" loop gain is still equal to unity.[2]

Example 14.2

Shown in Fig. 14.12 is a differential implementation of the oscillator of Fig. 14.8. What is the maximum voltage swing of each stage?

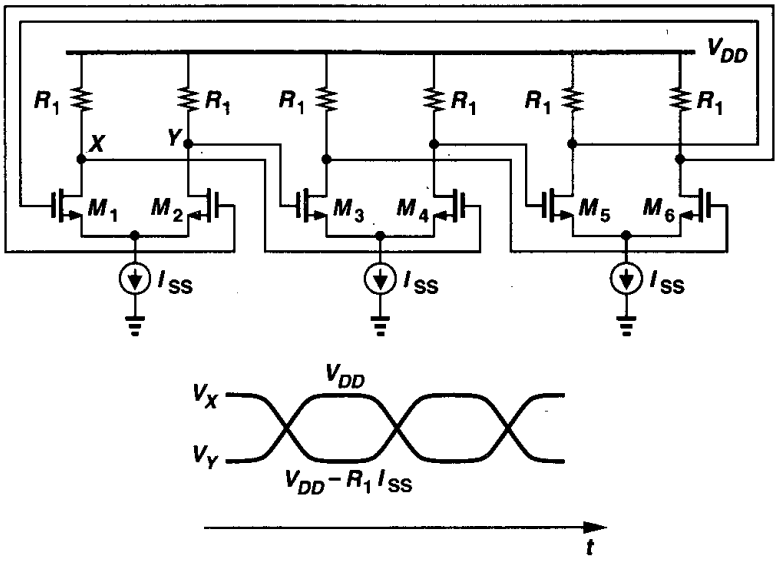


Figure 14.12

Solution

If the gain per stage is well above 2, then the amplitude grows until each differential pair experiences complete switching, that is, until I_{SS} is completely steered to one side every half cycle. As a result, the swing at each node is equal to $I_{SS}R_1$. From the waveforms shown in Fig. 14.12, we also observe that each stage is in its high-gain region for only a fraction of the period, (e.g., when $|V_X - V_Y|$ is small).

A simple implementation of ring oscillators that does not require resistors is depicted in Fig. 14.13. Suppose the circuit is released with an initial voltage at each node equal

[2]While intuitive, these statements are not rigorous. The concepts of transfer function, poles, and loop gain are difficult to apply to a nonlinear circuit.

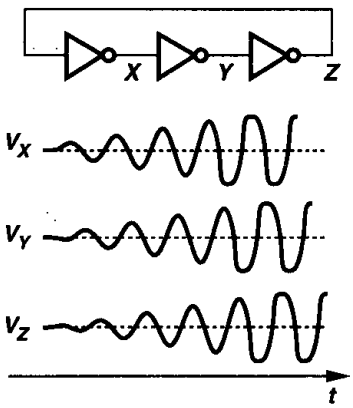


Figure 14.13 Ring oscillator using CMOS inverters.

to the trip point of the inverters, V_{trip}.[3] With identical stages and no noise in the devices, the circuit would remain in this state indefinitely,[4] but noise components disturb each node voltage, yielding a growing waveform. The signal eventually exhibits rail-to-rail swings.

Let us now assume the circuit of Fig. 14.13 begins with $V_X = V_{DD}$ (Fig. 14.14). Under this condition, $V_Y = 0$ and $V_Z = V_{DD}$. Thus, when the circuit is released, V_X begins to fall

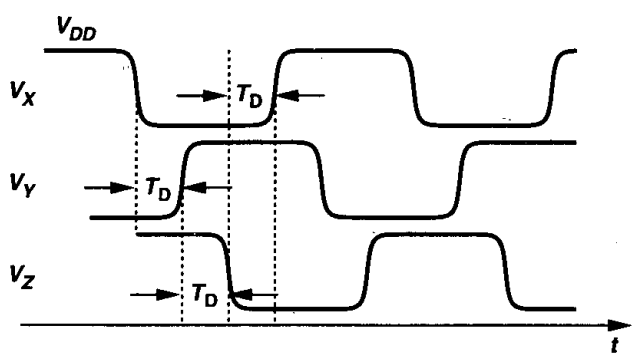


Figure 14.14 Waveforms of ring oscillator when one node is initialized at V_{DD}.

to zero (because the first inverter senses a high input), forcing V_Y to rise to V_{DD} after one inverter delay, T_D, and V_Z to fall to zero after another inverter delay. The circuit therefore oscillates with a delay of T_D between consecutive node voltages, yielding a period of $6T_D$.

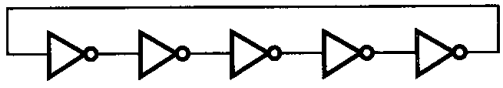
The above small-signal and large-signal analyses raise an interesting question. While the small-signal oscillation frequency is given by $A_0\sqrt{3}\omega_0/2$ [from Eq. (14.16)], the large-signal

[3]The trip point of an inverter is the input voltage that results in an equal output voltage.

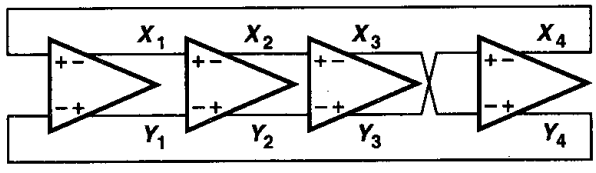
[4]This is indeed how SPICE predicts the circuit's behavior. To start the oscillation in SPICE, one of the nodes must be initialized at a different voltage.

value is $1/(6T_D)$. Are these two values equal? Not necessarily. After all, ω_0 is determined by the small-signal output resistance and capacitance of each inverter near the trip point whereas T_D results from the large-signal, nonlinear current drive and capacitances of each stage. In other words, when the circuit is released with all inverters at their trip point, the oscillation begins with a frequency of $\sqrt{3}A_0\omega_0/2$ but, as the amplitude grows and the circuit becomes nonlinear, the frequency shifts to $1/(6T_D)$ (which is a lower value).

Ring oscillators employing more than three stages are also feasible. The total number of inversions in the loop must be odd so that the circuit does not latch up. For example, as shown in Fig. 14.15(a), a ring can incorporate five inverters, providing a frequency of



(a)



(b)

Figure 14.15 (a) Five-stage single-ended ring oscillator, (b) four-stage differential ring oscillator.

$1/(10T_D)$. On the other hand, the differential implementation can utilize an *even* number of stages by simply configuring one stage such that it does not invert. Illustrated in Fig. 14.15(b), this flexibility demonstrates another advantage of differential circuits over their single-ended counterparts.

Example 14.3

What is the minimum required voltage gain per stage in the four-stage oscillator of Fig. 14.15(b)? How many signal phases are provided by the circuit?

Solution

Using a notation similar to that for Fig. 14.8, we have:

$$H(s) = -\frac{A_0^4}{(1+\frac{s}{\omega_0})^4}. \tag{14.17}$$

For the circuit to oscillate, each stage must contribute a frequency-dependent phase shift of $180°/4 = 45°$. The frequency at which this occurs is given by $\tan^{-1}\omega_{osc}/\omega_0 = 45°$ and hence $\omega_{osc} = \omega_0$. The

minimum voltage gain is therefore derived as

$$\frac{A_0}{\sqrt{1+(\frac{\omega_{osc}}{\omega_0})^2}} = 1. \tag{14.18}$$

That is, $A_0 = \sqrt{2}$. As expected, this value is lower than that required in a three-stage ring.

With 45° of phase shift per stage, the oscillator provides four phases and their complements. This is illustrated in Fig. 14.16.

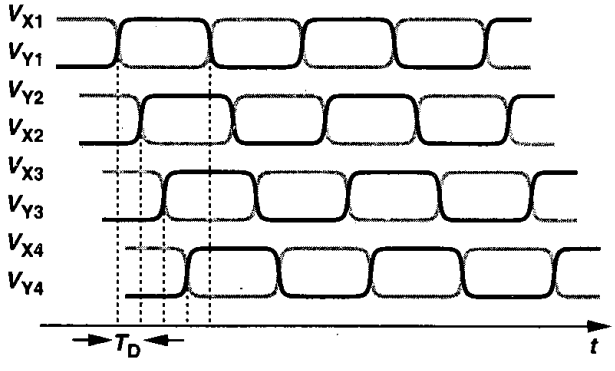


Figure 14.16

The number of stages in a ring oscillator is determined by various requirements, including speed, power dissipation, noise immunity, etc. In most applications, three to five stages provide optimum performance (for differential implementations).

Example 14.4

Determine the maximum voltage swings and the minimum supply voltage of a ring oscillator incorporating differential pairs with resistive loads (e.g., as in Fig. 14.12) if no transistor must enter the triode region. Assume each stage experiences complete switching.

Solution

Figure 14.17(a) shows two stages in cascade. If each stage experiences complete switching, then each drain voltage, e.g., V_X or V_Y, varies between V_{DD} and $V_{DD} - I_{SS}R_P$. Thus, when M_1 is fully on, its gate and drain voltages are equal to V_{DD} and $V_{DD} - I_{SS}R_P$, respectively. For this transistor to remain in saturation, we have $I_{SS}R_P \leq V_{TH}$, i.e., the peak-to-peak swing at each drain must not exceed V_{TH}.

How is the minimum supply voltage determined? If V_{DD} is lowered, the voltage at the common source node of each differential pair, e.g., V_P in Fig. 14.17(a), falls, eventually driving the tail transistor into the triode region. We must therefore calculate V_P for the worst case, noting that V_P does vary with time because M_1 and M_2 carry unequal currents when the input difference becomes large.

Now consider the stand-alone circuit of Fig. 14.17(b), assuming the inputs vary between V_{DD} and $V_{DD} - I_{SS}R_P$. How does V_P vary? When the gate voltage of M_1, V_1, is equal to V_{DD} and M_1 carries

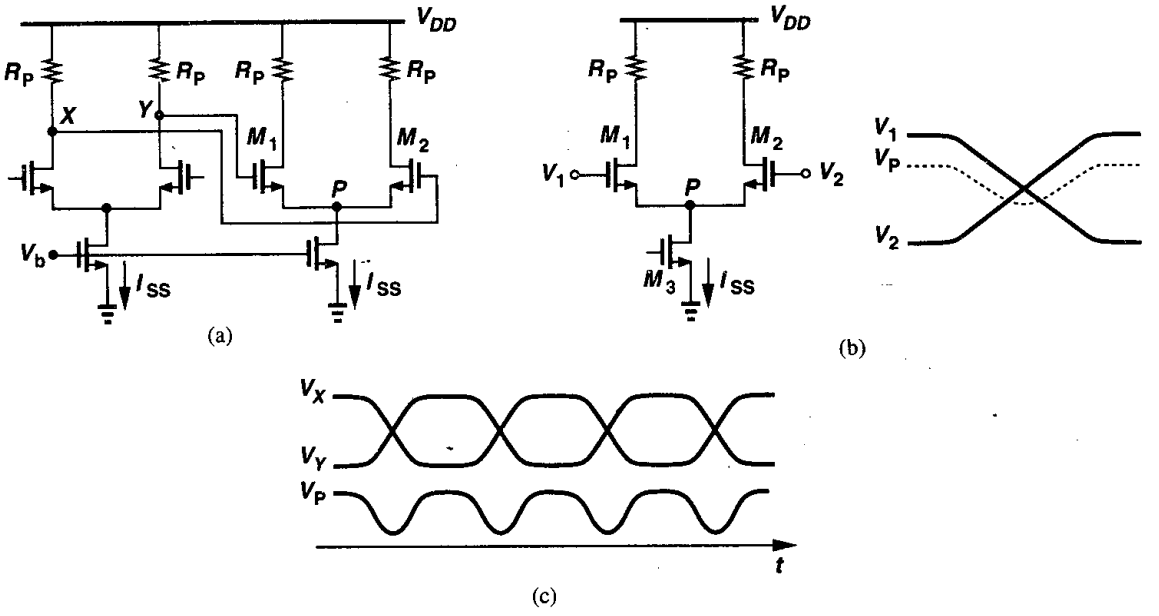


Figure 14.17

all of I_{SS},

$$V_P = V_{DD} - \sqrt{\frac{2I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} - V_{TH}. \tag{14.19}$$

As V_1 falls and V_2 rises, so does V_P because, so long as M_2 is off, M_1 operates as a source follower. When the difference between V_1 and V_2 reaches $\sqrt{2}(V_{GS,eq} - V_{TH})$, where $V_{GS,eq}$ denotes the equilibrium overdrive of each transistor, M_2 turns on. To calculate V_P after this point, we note that $I_{D1} + I_{D2} = I_{SS}$, $V_{GS1} = V_1 - V_P$, and $V_{GS2} = V_2 - V_P$. Thus,

$$\frac{1}{2}\mu_n C_{ox}(\frac{W}{L})_{1,2}(V_1 - V_P - V_{TH})^2 + \frac{1}{2}\mu_n C_{ox}(\frac{W}{L})_{1,2}(V_2 - V_P - V_{TH})^2 = I_{SS}. \tag{14.20}$$

Expanding the quadratic terms and rearranging the result, we have

$$2V_P^2 - 2(V_1 - V_{TH} + V_2 - V_{TH})V_P + (V_1 - V_{TH})^2 + (V_2 - V_{TH})^2 - \frac{2I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}} = 0. \tag{14.21}$$

It follows that

$$V_P = \frac{1}{2}[V_1 + V_2 - 2V_{TH} \pm \sqrt{-(V_1 - V_2)^2 + \frac{4I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}}]. \tag{14.22}$$

If V_1 and V_2 vary differentially, they can be expressed as $V_1 = V_{CM} + \Delta V$ and $V_2 = V_{CM} - \Delta V$, where $V_{CM} = V_{DD} - I_{SS}R_P/2$, yielding

$$V_P = V_{CM} - V_{TH} \pm \frac{1}{2}\sqrt{-(2\Delta V)^2 + \frac{4I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}}. \tag{14.23}$$

This expression reveals why node P is considered a virtual ground in small-signal operation: if $|\Delta V|$ is much less than the maximum overdrive voltage, then V_P is relatively constant. Since the term under the square root reaches a maximum for $\Delta V = 0$ (equilibrium condition),

$$V_{P,min} = V_{CM} - V_{TH} - \sqrt{\frac{I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}}. \tag{14.24}$$

As expected, the last term in (14.24) represents the overdrive voltage of each transistor in equilibrium (where $I_{D1} = I_{D2} = I_{SS}/2$).

Figure 14.17(c) shows typical waveforms in the oscillator. Note that V_P varies at twice the oscillation frequency. This property is sometimes exploited in "frequency doublers."

To determine the minimum supply voltage, we write $V_{P,min} \geq V_{ISS}$, where V_{ISS} denotes the minimum required voltage across I_{SS}. Thus,

$$V_{DD} - \frac{R_P I_{SS}}{2} - V_{TH} - \sqrt{\frac{I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} \geq V_{ISS}, \tag{14.25}$$

and

$$V_{DD} \geq V_{ISS} + V_{TH} + \sqrt{\frac{I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} + \frac{R_P I_{SS}}{2}. \tag{14.26}$$

The terms on the right are: the voltage headroom consumed by a current source, one threshold voltage, the equilibrium overdrive, and half of the swing at each node.

In CMOS technologies lacking high-quality resistors, the implementation of Fig. 14.17(a) must be modified. While a PMOS transistor operating in the deep triode region can serve as the load [Fig. 14.18(a)], the gate voltage must be set so as to define the on-resistance accurately. Alternatively, a diode-connected load can be utilized [Fig. 14.18(b)] but at the cost of one threshold voltage in the headroom. Figure 14.18(c) shows a more efficient load where an NMOS source follower is inserted between the drain and gate of each PMOS transistor. With the output sensed at nodes X and Y, M_3 and M_4 consume only a voltage headroom equal to $|V_{DS3,4}|$. If $V_{GS5} \approx V_{TH3}$, then M_3 operates at the edge of the triode region and the small-signal resistance of the load is roughly equal to $1/g_{m3}$ (with the assumption $\lambda = \gamma = 0$) (Problem 14.4).

The load of Fig. 14.18(c) exhibits another interesting property as well. Since the gate-source capacitance of M_3 is driven by the source follower, the time constant associated with the load is smaller than that of a diode-connected transistor. Also, the finite output resistance of the follower may yield an inductive behavior for the load (Problem 14.5).

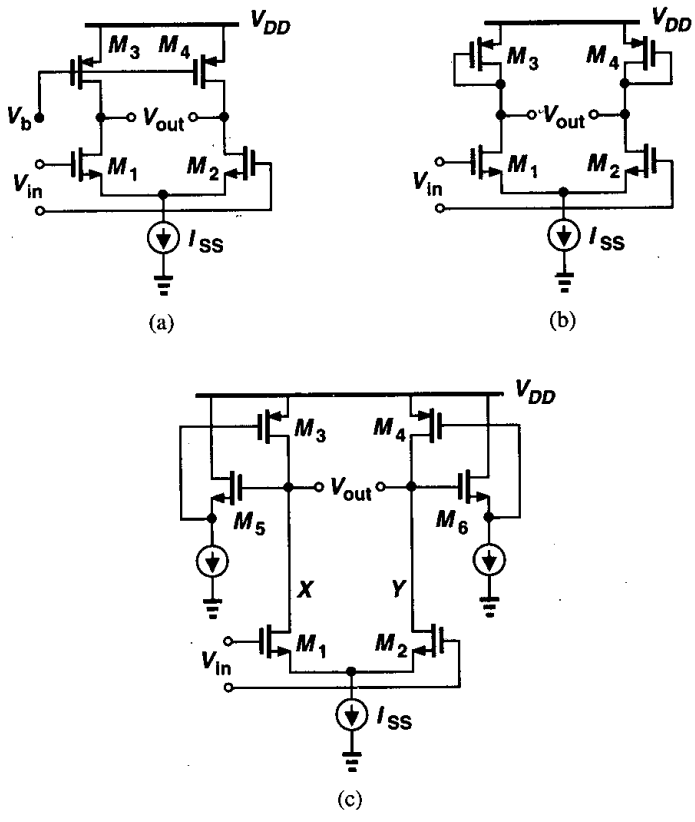


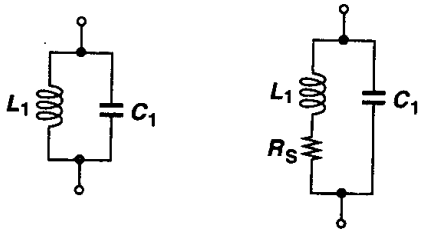
Figure 14.18 Differential stages using PMOS loads.

14.3 LC Oscillators

Monolithic inductors have gradually appeared in bipolar and CMOS technologies in the past 10 years, making it possible to design oscillators based on passive resonant circuits. Before delving into such oscillators, it is instructive to review basic properties of RLC circuits.

As shown in Fig. 14.19(a), an inductor L_1 placed in parallel with a capacitor C_1 resonates at a frequency $\omega_{res} = 1/\sqrt{L_1 C_1}$. At this frequency, the impedances of the inductor, $jL_1\omega_{res}$, and the capacitor, $1/(jC_1\omega_{res})$, are equal and opposite, thereby yielding an infinite impedance. We say the circuit has an infinite quality factor, Q. In practice, inductors (and capacitors) suffer from resistive components. For example, the series resistance of the metal wire used in the inductor can be modeled as shown in Fig. 14.19(b). We define the Q of the inductor as $L_1\omega/R_S$. For this circuit, the reader can show that the equivalent impedance is given by

$$Z_{eq}(s) = \frac{R_S + L_1 s}{1 + L_1 C_1 s^2 + R_S C_1 s}, \tag{14.27}$$



(a) (b)

Figure 14.19 (a) Ideal and (b) realistic LC tanks.

and hence,

$$|Z_{eq}(s = j\omega)|^2 = \frac{R_S^2 + L_1^2\omega^2}{(1 - L_1C_1\omega^2)^2 + R_S^2C_1^2\omega^2}. \tag{14.28}$$

That is, the impedance does not go to infinity at any $s = j\omega$. We say the circuit has a finite Q. The magnitude of Z_{eq} in (14.28) reaches a peak in the vicinity of $\omega = 1/\sqrt{L_1C_1}$, but the actual resonance frequency has some dependency on R_S.

The circuit of Fig. 14.19(b) can be transformed to an equivalent topology that more easily lends itself to analysis and design. To this end, we first consider the series combination shown in Fig. 14.20(a). For a narrow frequency range, it is possible to convert the circuit to the parallel configuration of Fig. 14.20(b). For the two impedances to be equivalent:

$$L_1s + R_S = \frac{R_PL_Ps}{R_P + L_Ps}. \tag{14.29}$$

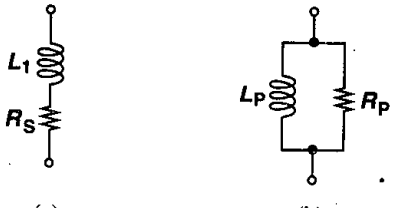
Considering only the steady state response, we assume $s = j\omega$ and rewrite (14.29) as

$$(L_1R_P + L_PR_S)j\omega + R_SR_P - L_1L_P\omega^2 = R_PL_Pj\omega. \tag{14.30}$$

This relationship must hold for all values of ω (in a narrow range), mandating that

$$L_1R_P + L_PR_S = R_PL_P \tag{14.31}$$

$$R_SR_P - L_1L_P\omega^2 = 0. \tag{14.32}$$



(a) (b)

Figure 14.20 Conversion of a series combination to a parallel combination.

Calculating R_P from the latter and substituting in the former, we have

$$L_P = L_1(1 + \frac{R_S^2}{L_1^2\omega^2}). \tag{14.33}$$

Recall that $L_1\omega/R_S = Q$, a value typically greater than 3 for monolithic inductors. Thus,

$$L_P \approx L_1 \tag{14.34}$$

and

$$R_P \approx \frac{L_1^2\omega^2}{R_S} \tag{14.35}$$

$$\approx Q^2 R_S. \tag{14.36}$$

In other words, the parallel network has the same reactance but a resistance Q^2 times the series resistance. This concept holds valid for a first-order RC network as well if the Q of the series combination is defined as $1/(C\omega)/R_S$.

The above transformation allows the conversion illustrated in Fig. 14.21, where $C_P = C_1$. The equivalence of course breaks down as ω departs susbtantially from the resonance

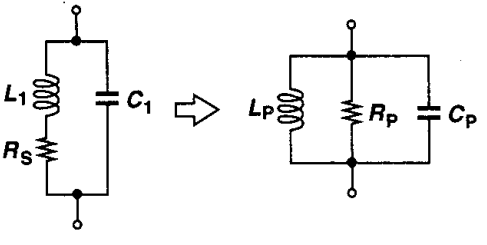


Figure 14.21 Conversion of a tank to three parallel components.

frequency. The insight gained from the parallel combination is that at $\omega_1 = 1/\sqrt{L_pC_p}$, the tank reduces to a simple resistor; i.e., the phase difference between the voltage and current of the tank drops to zero. Plotting the magnitude of the tank impedance versus frequency [Fig. 14.22(a)], we note that the behavior is inductive for $\omega < \omega_1$ and capacitive for $\omega > \omega_1$. We then surmise that the phase of the impedance is positive for $\omega < \omega_1$ and negative for $\omega > \omega_1$ [Fig. 14.22(b)]. These observations prove useful in studying LC oscillators. (Why do we expect the phase shift to approach $+90°$ at very low frequencies and $-90°$ at very high frequencies?)

Let us now consider the "tuned" stage of Fig. 14.23(a), where an LC tank operates as the load. At resonance, $jL_p\omega = 1/(jC_p\omega)$ and the voltage gain equals $-g_{m1}R_P$. (Note that the gain of the circuit is very small at frequencies near zero.) Does this circuit oscillate if the output is connected to the input [Fig. 14.23(b)]? At resonance, the total phase shift around the loop is equal to 180° (rather than 360°). Also, from Fig. 14.22(b), the frequency-dependent phase shift of the tank never reaches 180°. Thus, the circuit does not oscillate.

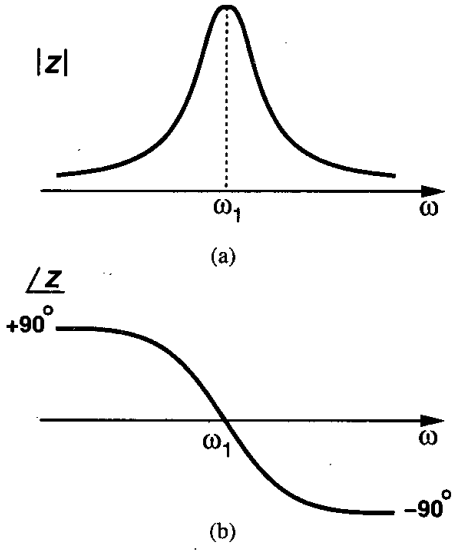


Figure 14.22 (a) Magnitude and (b) phase of the impedance of an LC tank as a function of frequency.

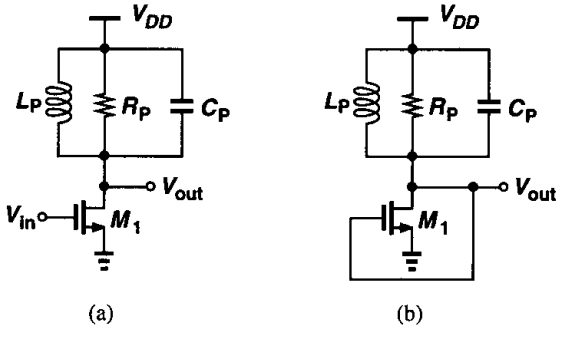


Figure 14.23 (a) Tuned gain stage, (b) stage of (a) in feedback.

Before modifying the circuit for oscillatory behavior, let us observe another interesting property of the gain stage of Fig. 14.23(a) that distinguishes it from a common-source topology using a resistive load. Suppose, as shown in Fig. 14.24, the stage is biased at a drain current I_1. If the series resistance of L_p is small, the dc level of V_{out} is close to V_{DD}. How does V_{out} vary if a small sinusoidal voltage at the resonance frequency is applied to the input? We expect V_{out} to be an inverted sinusoid with an average value near V_{DD} because the inductor cannot sustain a large dc drop. In other words, if the average value of V_{out} deviates significantly from V_{DD}, then the inductor series resistance must carry an average current greater than I_1. Thus, the peak output level in fact *exceeds* the supply voltage, an

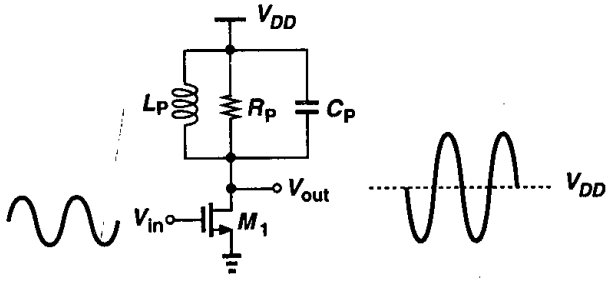


Figure 14.24 Output signal levels in a tuned stage.

important and often useful attribute of the LC load. For example, with proper design, the output peak-to-peak swing can be larger than V_{DD}.

We now study two types of LC oscillators.

14.3.1 Crossed-Coupled Oscillator

Suppose we place two stages of Fig. 14.23(a) in a cascade, as depicted in Fig. 14.25. While similar to the topology of Fig. 14.5, this configuration does not latch up because its low-

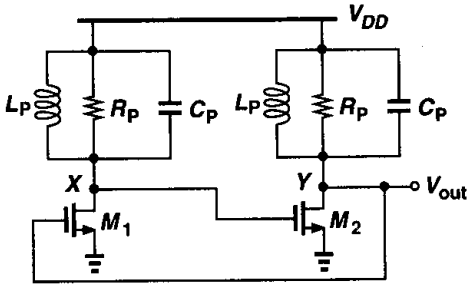


Figure 14.25 Two tuned stages in a feedback loop.

frequency gain is very small. Furthermore, at resonance, the total phase shift around the loop is zero because each stage contributes zero frequency-dependent phase shift. That is, if $g_{m1}R_P g_{m2}R_P \geq 1$, then the loop oscillates. Note that V_X and V_Y are differential waveforms. (Why?)

Example 14.5

Sketch the open-loop voltage gain and phase of the circuit shown in Fig. 14.25. Neglect transistor capacitances.

Solution

The magnitude of the transfer function has a shape similar to that in Fig. 14.22(a) but with sharper rise and fall because it results from the *product* of those of the two stages. The total phase at low frequencies is given by signal inversion by each common-source stage plus a 90° phase shift due to each tank.

A similar behavior occurs at high frequencies. The gain and phase are sketched in Fig. 14.26. From these plots, the reader can prove that the circuit cannot oscillate at any other frequency.

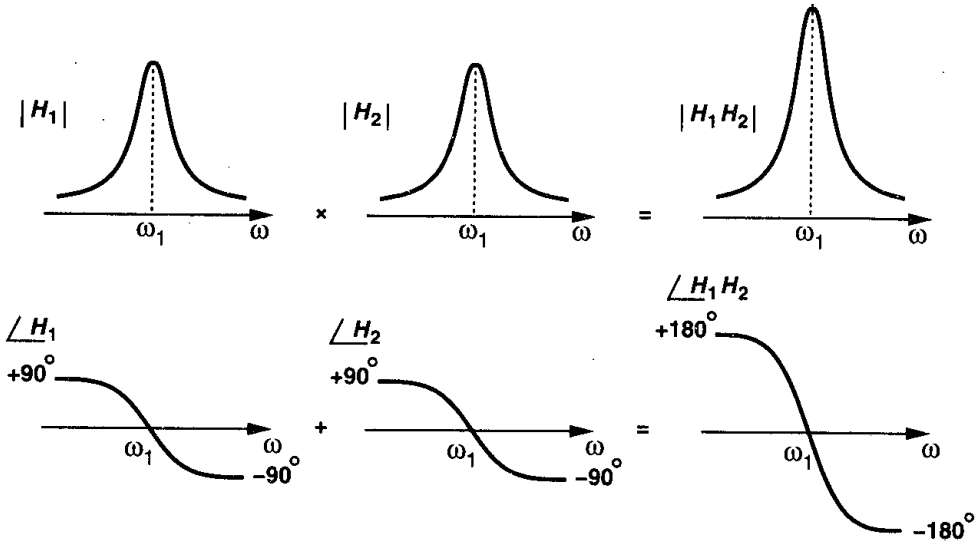


Figure 14.26 Loop gain characteristics of the circuit shown in Fig. 14.25.

The circuit of Fig. 14.25 serves as the core of many LC oscillators and is sometimes drawn as in Fig. 14.27(a) or (b). However, the drain currents of M_1 and M_2 and hence the output swings heavily depend on the supply voltage. Since the waveforms at X and Y are differential, the drawing in Fig. 14.27(b) suggests that M_1 and M_2 can be converted to a differential pair as depicted in Fig. 14.27(c), where the total bias current is defined by I_{SS}.

Example 14.6

For the circuit of Fig. 14.27(c), plot V_X and V_Y and I_{D1} and I_{D2} as the oscillation begins.

Solution

If the circuit begins with zero difference between V_X and V_Y, then $V_X = V_Y \approx V_{DD}$. The two transistors share the tail current equally. If $(g_{m1,2}R_P)^2 \geq 1$, where R_P is the equivalent parallel resistance of the tank at resonance, then noise components at the resonance frequency are continually amplified by M_1 and M_2, allowing the oscillation to grow. The drain currents of M_1 and M_2 vary according to the instantaneous value of $V_X - V_Y$ (as in a differential pair).

As shown in Fig. 14.28, the oscillation amplitude grows until the loop gain drops at the peaks. In fact, if $g_{m1,2}R_P$ is large enough, the difference between $V_X - V_Y$ reaches a level that steers the entire tail current to one transistor, turning the other off. Thus, in the steady state, I_{D1} and I_{D2} vary between zero and I_{SS}.

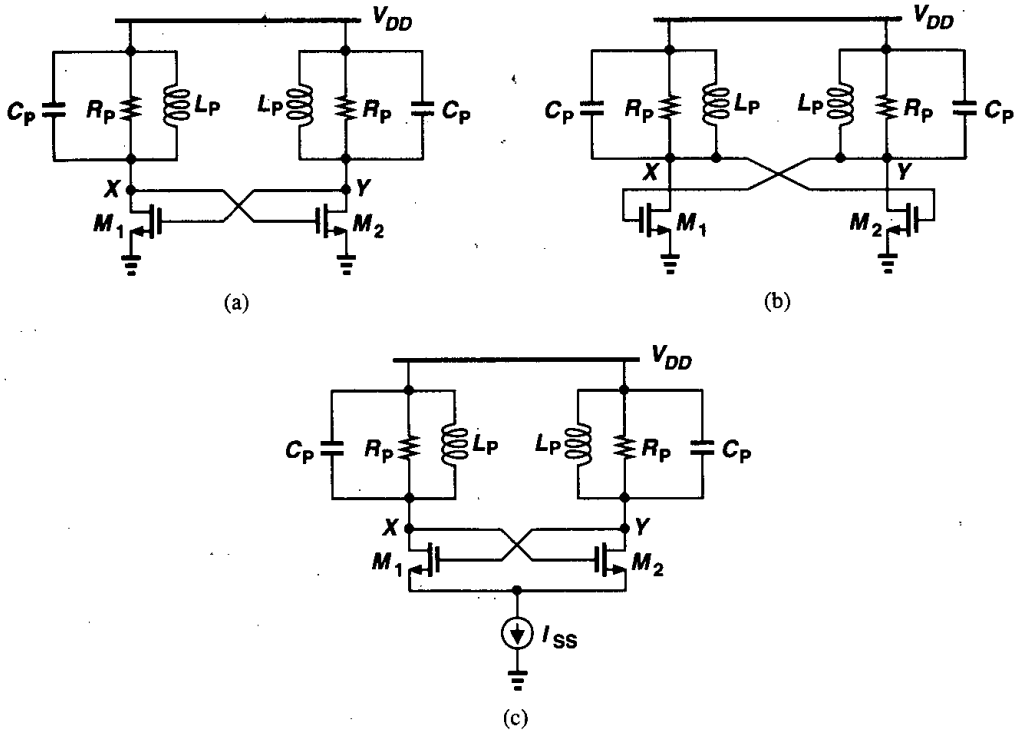


Figure 14.27 (a) Redrawing of the oscillator shown in Fig. 14.25, (b) another redrawing of the circuit, (c) addition of tail current source to lower supply sensitivity.

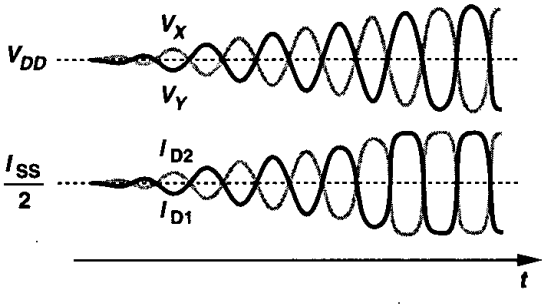


Figure 14.28

The oscillator of Fig. 14.27(c) is constructed in fully differential form. The supply sensitivity of the circuit, however, is nonzero even with perfect symmetry. This is because the drain junction capacitances of M_1 and M_2 vary with the supply voltage. We return to this issue in Example 14.9.

14.3.2 Colpitts Oscillator

An LC oscillator may be realized with only one transistor in the signal path. Consider the gain stage of Fig. 14.23(a) again and recall that the drain voltage cannot be applied to the gate because the overall phase shift at resonance equals 180° rather than 360°. Also, recall that in a common-gate stage, the phase shift from the source to the drain is zero. We then surmise that if, as shown in Fig. 14.29(a), the drain voltage is returned to the source rather than the gate, the circuit may oscillate. The coupling must incorporate a capacitor to avoid disturbing the bias point of M_1.

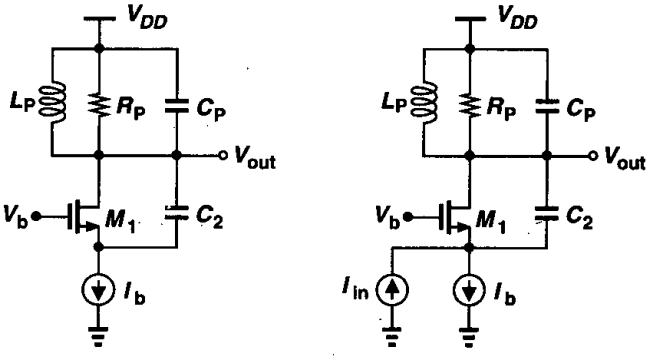


Figure 14.29 (a) Tuned stage with feedback applied from drain to source, (b) addition of input current to calculate closed-loop gain.

Unfortunately, owing to insufficient loop gain, the circuit of Fig. 14.29(a) does not oscillate. To prove this point, we invoke the view of Fig. 14.1, where an oscillator is considered a feedback system with infinite closed-loop gain. Applying an input current as depicted in Fig. 14.29(b) and neglecting transistor parasitics, we obtain the closed-loop gain as:

$$\frac{V_{out}}{I_{in}} = L_{PS} || \frac{1}{C_{PS}} || R_P \tag{14.37}$$

because M_1 and C_2 directly conduct the input current to the tank. Since the closed-loop gain cannot be equal to infinity at any frequency, the circuit fails to oscillate.

Example 14.7

The reader may wonder why the input to the feedback system is realized as a current source applied to the source of the transistor rather than a voltage source applied to its gate. Perform the analysis with the latter stimulus.

Solution

From Fig. 14.30, we note that with a finite variation of V_{in}, the change in I_b is still zero if the bias current source is ideal. Thus, if the source-bulk junction capacitance of M_1 is neglected, the

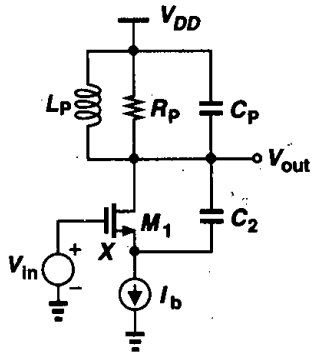


Figure 14.30

change in the tank current is zero, yielding $V_{out}/V_{in} = 0$. Interestingly, V_X does vary with V_{in}, but M_1 generates a small-signal current that cancels that through C_2. The reader can prove that $V_X/V_{in} = g_m/(g_m + C_2 s)$.

The above example reveals two important points. First, to excite a circuit into oscillation, the stimulus can be applied at different points. (That is, the noise of any device in the loop can initiate the oscillation.[5]) Second, in Fig. 14.30, V_{out}/V_{in} is zero because the impedance connected between the source of M_1 and ground is infinity. We then add a capacitor from this node to ground as shown in Fig. 14.31(a), seeking conditions of oscillation. Note that the capacitor in parallel with L_P is removed. The reason will become clear later.

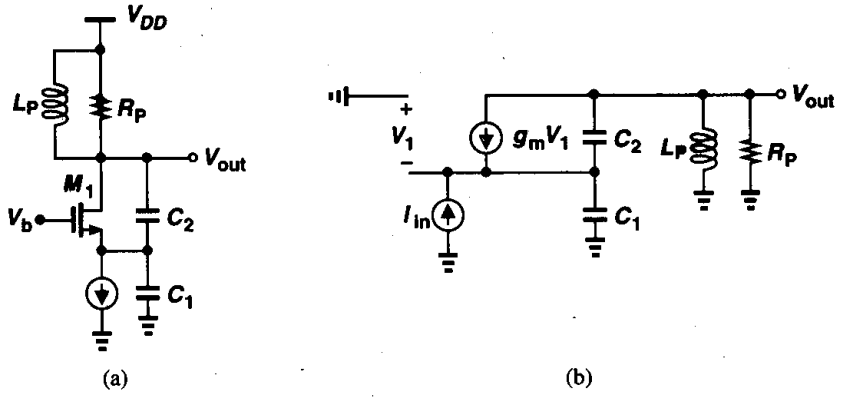


Figure 14.31 (a) Colpitts oscillator, (b) equivalent circuit of (a) with input stimulus.

[5]This is because the natural frequencies of a linear (observable) system do not depend on the location of the stimulus. Of course, the type of stimulus (voltage or current) must be chosen such that when it is set to zero, the circuit returns to its original topology. For example, driving the gate of M_1 in Fig. 14.30 by a current changes the natural frequencies of the circuit.

Approximating M_1 by a single voltage-dependent current source, we construct the equivalent circuit of Fig. 14.31(b). Since the current through the parallel combination of L_P and R_P is given by $V_{out}/(L_P s) + V_{out}/R_P$, the total current through C_1 is equal to $I_{in} - V_{out}/(L_P s) - V_{out}/R_P$, yielding

$$V_1 = -(I_{in} - \frac{V_{out}}{L_P s} - \frac{V_{out}}{R_P})\frac{1}{C_1 s}. \tag{14.38}$$

Writing the current through C_2 as $(V_{out} + V_1)C_2 s$, we sum all of the currents at the output node:

$$-g_m(I_{in} - \frac{V_{out}}{L_P s} - \frac{V_{out}}{R_P})\frac{1}{C_1 s} + [V_{out} - (I_{in} - \frac{V_{out}}{L_P s} - \frac{V_{out}}{R_P})\frac{1}{C_1 s}]C_2 s + \frac{V_{out}}{L_P s} + \frac{V_{out}}{R_P} = 0. \tag{14.39}$$

It follows that

$$\frac{V_{out}}{I_{in}} = \frac{R_P L_P s(g_m + C_2 s)}{R_P C_1 C_2 L_P s^3 + (C_1 + C_2)L_P s^2 + [g_m L_P + R_P(C_1 + C_2)]s + g_m R_P}. \tag{14.40}$$

Note that, as expected, (14.40) reduces to $(L_P s \| R_P)$ if $C_1 = 0$. The circuit oscillates if the closed-loop transfer function goes to infinity at an imaginary value of s, $s_R = j\omega_R$. Consequently, both the real and imaginary parts of the denominator must drop to zero at this frequency:

$$-R_P C_1 C_2 L_P \omega_R^3 + [g_m L_P + R_P(C_1 + C_2)]\omega_R = 0 \tag{14.41}$$

$$-(C_1 + C_2)L_P \omega_R^2 + g_m R_P = 0. \tag{14.42}$$

Since with typical values, $g_m L_P \ll R_P(C_1 + C_2)$, Eq. (14.41) yields:

$$\omega_R^2 = \frac{1}{L_P \frac{C_1 C_2}{C_1 + C_2}}, \tag{14.43}$$

and Eq. (14.42) results in

$$g_m R_P = \frac{(C_1 + C_2)^2}{C_1 C_2} \tag{14.44}$$

$$= \frac{C_1}{C_2}(1 + \frac{C_2}{C_1})^2. \tag{14.45}$$

Recognizing that $g_m R_P$ is the voltage gain from the source of M_1 to the output (if $g_{mb} = 0$), we determine the ratio C_1/C_2 for minimum required gain. The reader can prove that the minimum occurs for $C_1/C_2 = 1$, requiring

$$g_m R_P \geq 4. \tag{14.46}$$

Equation (14.46) demonstrates an important disadvantage of the Colpitts oscillator with respect to the cross-coupled topology of Fig. 14.27(c). The former demands a voltage gain of at least 4 at resonance and the latter, only unity. This issue is critical if the inductor

suffers from a low Q and hence a small R_P, a common situation in CMOS technologies. As a consequence, the cross-coupled scheme is used more widely.

The foregoing analysis neglected the capacitance that appears in parallel with the inductor. As suggested in Problem 14.10, if this capacitance, C_P, is included in the equivalent circuit, Eq. (14.43) is modified as:

$$\omega_R^2 = \frac{1}{L_P(C_P + \dfrac{C_1 C_2}{C_1 + C_2})}, \tag{14.47}$$

whereas (14.46) remains unchanged. Thus, C_P is simply included in parallel with the series combination of C_1 and C_2.

14.3.3 One-Port Oscillators

Our development of oscillators thus far has been based on feedback systems. An alternative view that provides more insight into the oscillation phenomenon employs the concept of "negative resistance." To arrive at this view, let us first consider a simple tank that is stimulated by a current impulse [Fig. 14.32(a)]. The tank responds with a decaying oscillatory behavior because, in every cycle, some of the energy that reciprocates between

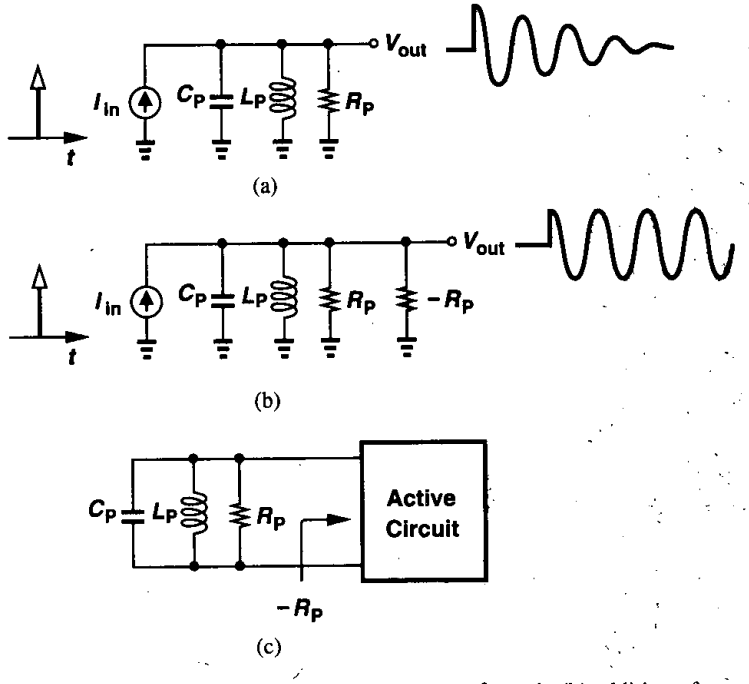


Figure 14.32 (a) Decaying impulse response of a tank, (b) addition of negative resistance to cancel loss in R_P, (c) use of an active circuit to provide negative resistance.

the capacitor and the inductor is lost in the form of heat in the resistor. Now suppose a resistor equal to $-R_P$ is placed in parallel with R_P and the experiment is repeated [Fig. 14.32(b)]. Since $R_P||(-R_P) = \infty$, the tank oscillates indefinitely. Thus, if a one-port circuit exhibiting a negative resistance is placed in parallel with a tank [Fig. 14.32(c)], the combination may oscillate. Such a topology is called a one-port oscillator.

How can a circuit provide a negative resistance? Recall that feedback multiplies or divides the input and output impedances of circuits by a factor equal to one plus the loop gain. Thus, if the loop gain is sufficiently *negative*, (i.e., the feedback is sufficiently positive), a negative resistance is achieved. As a simple example, let us apply positive feedback around a source follower. The follower introduces no signal inversion and neither must the feedback network. As depicted in Fig. 14.33(a), we implement the feedback by a common-gate stage

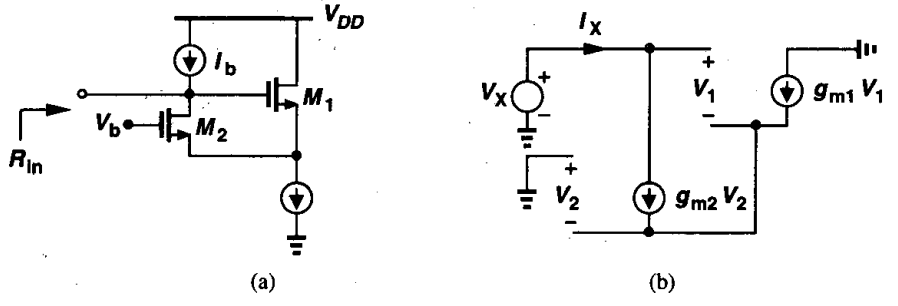


Figure 14.33 (a) Source follower with positive feedback to create negative input impedance, (b) equivalent circuit of (a) to calculate the input impedance.

and add the current source I_b to provide the bias current of M_2. From the equivalent circuit in Fig. 14.33(b) (where channel-length modulation and body effect are neglected), we have

$$I_X = g_{m2}V_2 = -g_{m1}V_1 \tag{14.48}$$

and

$$V_X = V_1 - V_2 \tag{14.49}$$

$$= -\frac{I_X}{g_{m1}} - \frac{I_X}{g_{m2}}. \tag{14.50}$$

Thus,

$$\frac{V_X}{I_X} = -(\frac{1}{g_{m1}} + \frac{1}{g_{m2}}), \tag{14.51}$$

and, if $g_{m1} = g_{m2} = g_m$, then

$$\frac{V_X}{I_X} = \frac{-2}{g_m}. \tag{14.52}$$

Negative resistance becomes more intuitive if we bear in mind that it is an *incremental* quantity, that is, negative resistance indicates that if the applied voltage *increases*, the current drawn by the circuit *decreases*. In Fig. 14.33(a), for example, if the input voltage increases, so does the source voltage of M_1, decreasing the drain current of M_2 and allowing part of I_b to flow to the input source.

With a negative resistance available, we can now construct an oscillator as illustrated in Fig. 14.34. Here, R_P denotes the equivalent parallel resistance of the tank and, for oscillation

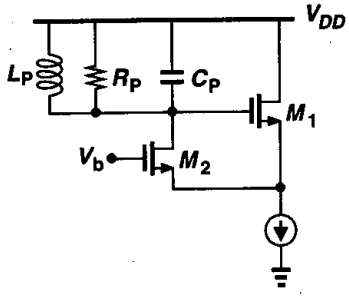


Figure 14.34 Oscillator using negative input resistance of a source follower with positive feedback.

build-up, $R_P - 2/g_m \geq 0$. Note that the inductor provides the bias current of M_2, obviating the need for a current source. If the small-signal resistance presented by M_1 and M_2 to the tank is less negative than $-R_P$, then the circuit experiences large swings such that each transistor is nearly off for part of the period, thereby yielding an "average" resistance of $-R_P$.

The circuit of Fig. 14.34 is similar to the stage of Fig. 14.29(a) but with the feedback capacitor replaced by a source follower. More interestingly, the circuit can be redrawn as in Fig. 14.35(a), bearing a resemblance to Fig. 14.27(c). In fact, if the drain current of M_1 flows

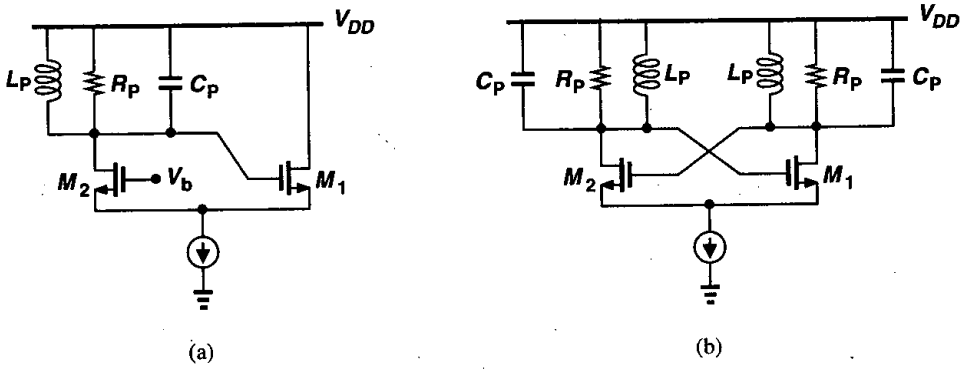


Figure 14.35 (a) Redrawing of the topology shown in Fig. 14.34, (b) differential version of (a).

through a tank and the resulting voltage is applied to the gate of M_2, the topology of Fig. 14.35(b) is obtained. Ignoring bias paths and merging the two tanks into one (Fig. 14.36), we note that the cross-coupled pair must provide a negative resistance of $-R_P$ between

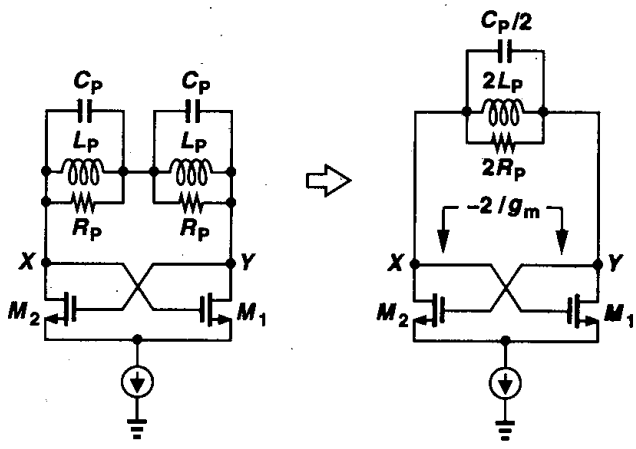


Figure 14.36 Equivalent circuit of Fig. 14.35(b).

nodes X and Y to enable oscillation. The reader can prove that this resistance is equal to $-2/g_m$ and hence it is necessary that $R_P \geq 1/g_m$. Thus, the circuit can be viewed as either a feedback system or a negative resistance in parallel with a lossy tank. This topology is also called a "negative-G_m oscillator."

As another method of creating negative resistance, consider the topology depicted in Fig. 14.37(a), where none of the nodes is grounded and channel-length modulation, body effect, and transistor capacitances are neglected. Since the drain current of M_1 is equal to $(-I_X/C_1 s)g_m$, we have

$$V_X = (I_X - \frac{-I_X}{C_1 s} g_m)\frac{1}{C_2 s} + \frac{I_X}{C_1 s} \tag{14.53}$$

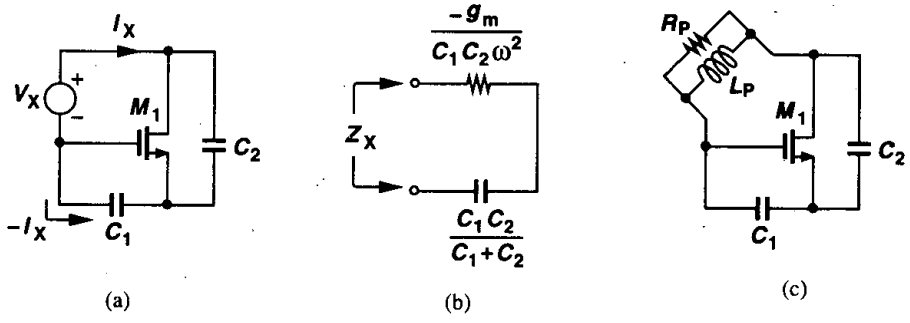


Figure 14.37 (a) Circuit topology providing negative resistance, (b) equivalent circuit of (a), (c) oscillator using (a).

and hence

$$\frac{V_X}{I_X} = \frac{g_m}{C_1 C_2 s^2} + \frac{1}{C_2 s} + \frac{1}{C_1 s}. \tag{14.54}$$

For $s = j\omega$, this impedance consists of a negative resistance equal to $-g_m/(C_1 C_2 \omega^2)$ in series with the series combination of C_1 and C_2 [Fig. 14.37(b)]. Thus, as shown in Fig. 14.37(c), if an inductor is placed between the gate and drain of M_1, the circuit may oscillate. Of the three nodes in the circuit, one can be an ac ground, resulting in the three different topologies illustrated in Fig. 14.38. The circuit of Fig. 14.38(a) is in fact based on a source follower, whose input impedance was found in Chapter 6 to contain a negative real part. The configuration of Fig. 14.38(b) is a Colpitts oscillator.

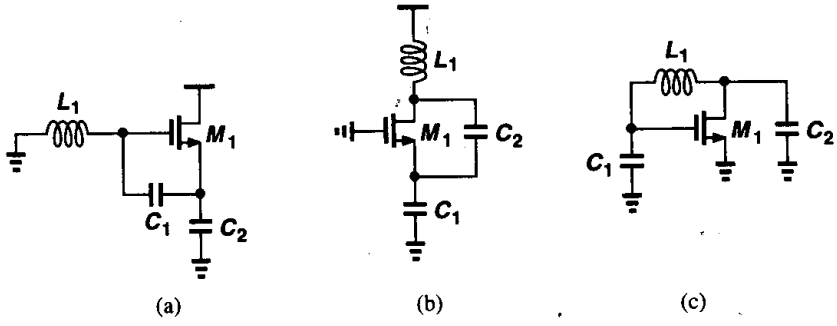


Figure 14.38 Oscillator topologies derived from the circuit of Fig. 14.37(c).

Example 14.8

Redraw the circuits of Fig. 14.38 with proper biasing.

Solution

The circuits are redrawn in Fig. 14.39.

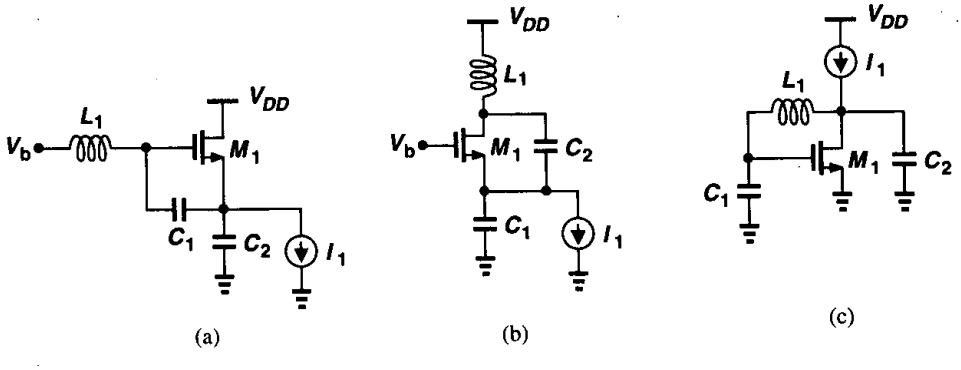


Figure 14.39

14.4 Voltage-Controlled Oscillators

Most applications require that oscillators be "tunable," i.e., their output frequency be a function of a control input, usually a voltage. An ideal voltage-controlled oscillator is a circuit whose output frequency is a linear function of its control voltage (Fig. 14.40):

$$\omega_{out} = \omega_0 + K_{VCO} V_{cont}. \tag{14.55}$$

Here, ω_0 represents the intercept corresponding to $V_{cont} = 0$ and K_{VCO} denotes the "gain" or "sensitivity" of the circuit (expressed in rad/s/V).[6] The achievable range, $\omega_2 - \omega_1$, is called the "tuning range."

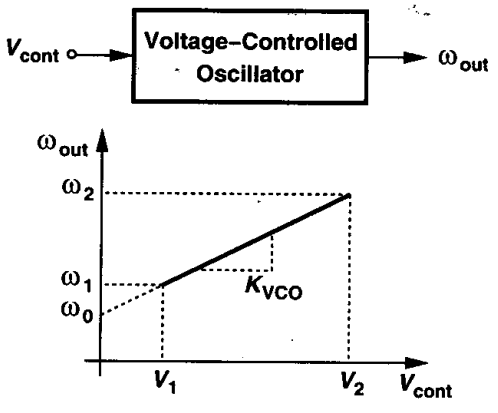


Figure 14.40 Definition of a VCO.

Example 14.9

In the negative-G_m oscillator of Fig. 14.27(c), assume $C_P = 0$, consider only the drain junction capacitance, C_{DB}, of M_1 and M_2, and explain why V_{DD} can be viewed as the control voltage. Calculate the gain of the VCO.

Solution

Since C_{DB} varies with the drain-bulk voltage, if V_{DD} changes, so does the resonance frequency of the tank. Noting that the average voltage across C_{DB} is approximately equal to V_{DD}, we write

$$C_{DB} = \frac{C_{DB0}}{(1 + \frac{V_{DD}}{\phi_B})^m}, \tag{14.56}$$

and

$$K_{VCO} = \frac{\partial \omega_{out}}{\partial V_{DD}} \tag{14.57}$$

$$= \frac{\partial \omega_{out}}{\partial C_{DB}} \cdot \frac{\partial C_{DB}}{\partial V_{DD}}. \tag{14.58}$$

[6] A more familiar unit is Hz/V but one must be careful with the dimension of K_{VCO} in the context of phase-locked loops.

With $\omega_{out} = 1/\sqrt{L_P C_{DB}}$, we have

$$K_{VCO} = \frac{-1}{2\sqrt{L_P C_{DB}} C_{DB}} \cdot \frac{-mC_{DB}}{\phi_B(1 + \frac{V_{DD}}{\phi_B})} \tag{14.59}$$

$$= \frac{m}{2\phi_B(1 + \frac{V_{DD}}{\phi_B})} \cdot \omega_{out}. \tag{14.60}$$

Note that the relationship between ω_{out} and V_{cont} is nonlinear because K_{VCO} varies with V_{DD} and ω_{out}.

Before modifying the oscillators studied in the previous sections for tunability, we summarize the important performance parameters of VCOs.

Center Frequency The center frequency (i.e., the midrange value in Fig. 14.40) is determined by the environment in which the VCO is used. For example, in the clock generation network of a microprocessor, the VCO may be required to run at the clock rate or even twice that. Today's CMOS VCOs achieve center frequencies as high as 10 GHz.

Tuning Range The required tuning range is dictated by two parameters: (1) the variation of the VCO center frequency with process and temperature and (2) the frequency range necessary for the application. The center frequency of some CMOS oscillators may vary by a factor of two at the extremes of process and temperature, thus mandating a sufficiently wide ($\geq 2\times$) tuning range to guarantee that the VCO output frequency can be driven to the desired value. Also, some applications incorporate clock frequencies that must vary by one to two orders of magnitude depending on the mode of operation, demanding a proportionally wide tuning range.

An important concern in the design of VCOs is the variation of the output phase and frequency as a result of noise on the control line. For a given noise amplitude, the noise in the output frequency is proportional to K_{VCO} because $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$. Thus, to minimize the effect of noise in V_{cont}, the VCO gain must be *minimized*, a constraint in direct conflict with the required tuning range. In fact, if, as shown in Fig. 14.40, the allowable range of V_{cont} is from V_1 to V_2 (e.g., from 0 to V_{DD}) and the tuning range must span at least ω_1 to ω_2, then K_{VCO} must satisfy the following requirement:

$$K_{VCO} \geq \frac{\omega_2 - \omega_1}{V_2 - V_1}. \tag{14.61}$$

Note that, for a given tuning range, K_{VCO} increases as the supply voltage decreases, making the oscillator more sensitive to noise on the control line.

Tuning Linearity As exemplified by Eq. (14.60), the tuning characteristics of VCOs exhibit nonlinearity, i.e., their gain, K_{VCO}, is not constant. As explained in Chapter 15, such nonlinearity degrades the settling behavior of phase-locked loops. For this reason, it is desirable to minimize the variation of K_{VCO} across the tuning range.

Actual oscillator characteristics typically exhibit a high gain region in the middle of the range and a low gain at the two extremes (Fig. 14.41). Compared to a linear characteristic (the gray line), the actual behavior displays a maximum gain *greater* than that predicted

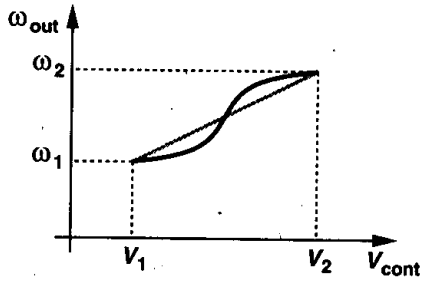


Figure 14.41 Nonlinear VCO characteristic.

by (14.61), implying that, for a given tuning range, nonlinearity inevitably leads to higher sensitivity for some region of the characteristic.

Output Amplitude It is desirable to achieve a large output oscillation amplitude, thus making the waveform less sensitive to noise. The amplitude trades with power dissipation, supply voltage, and (as explained in Section 14.4.2) even the tuning range. Also, the amplitude may vary across the tuning range, an undesirable effect.

Power Dissipation As with other analog circuits, oscillators suffer from trade-offs between speed, power dissipation, and noise. Typical oscillators drain 1 to 10 mW of power.

Supply and Common-Mode Rejection Oscillators are quite sensitive to noise, especially if they are realized in single-ended form. As seen in Example 14.9, even differential oscillators exhibit supply sensitivity. The design of oscillators for high noise immunity is a difficult challenge. Note that noise may be coupled to the control line of a VCO as well. For these reasons, it is preferable (but not always possible) to employ differential paths for both the oscillation signal and the control line.

Output Signal Purity Even with a constant control voltage, the output waveform of a VCO is not perfectly periodic. The electronic noise of the devices in the oscillator and supply noise lead to noise in the output phase and frequency. These effects are quantified by "jitter" and "phase noise" and determined by the requirements of each application.

14.4.1 Tuning in Ring Oscillators

Recall from Section 14.2 that the oscillation frequency, f_{osc}, of an N-stage ring equals $(2NT_D)^{-1}$, where T_D denotes the large-signal delay of each stage. Thus, to vary the frequency, T_D can be adjusted.

As a simple example, consider the differential pair of Fig. 14.42 as one stage of a ring oscillator. Here, M_3 and M_4 operate in the triode region, each acting as a variable resistor controlled by V_{cont}. As V_{cont} becomes more positive, the on-resistance of M_3 and M_4 increases, thus raising the time constant at the output, τ_1, and lowering f_{osc}. If M_3 and M_4

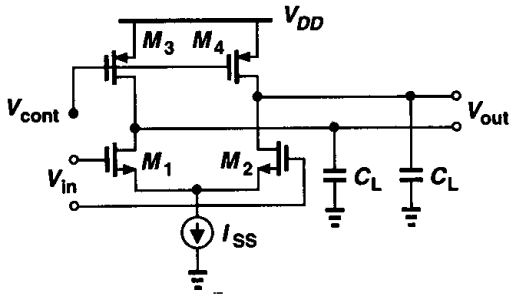


Figure 14.42 Differential pair with variable output time constant.

remain in deep triode region,

$$\tau_1 = R_{on3,4}C_L \tag{14.62}$$

$$= \frac{C_L}{\mu_p C_{ox}(\frac{W}{L})_{3,4}(V_{DD} - V_{cont} - |V_{THP}|)}. \tag{14.63}$$

In the above equation, C_L denotes the total capacitance seen at each output to ground (including the input capacitance of the following stage). The delay of the circuit is roughly proportional to τ_1, yielding

$$f_{osc} \propto \frac{1}{T_D} \tag{14.64}$$

$$\propto \frac{\mu_p C_{ox}(\frac{W}{L})_{3,4}(V_{DD} - V_{cont} - |V_{THP}|)}{C_L}. \tag{14.65}$$

Interestingly, f_{osc} is linearly proportional to V_{cont}.

Example 14.10

For given device dimensions and bias currents in Fig. 14.42, determine the maximum allowable value of V_{cont}. What happens if M_3 and M_4 enter saturation?

Solution

Let us assume (somewhat arbitrarily) that M_3 and M_4 remain in deep triode region if $|V_{DS3,4}| \leq 0.2 \times 2|V_{GS3,4} - V_{THP}|$. If each stage in the ring experiences complete switching, then the maximum drain current of M_3 and M_4 is equal to I_{SS}. To satisfy the above condition, we must have $I_{SS}R_{on3,4} \leq 0.4(V_{DD} - V_{cont} - |V_{THP}|)$, and hence

$$\frac{I_{SS}}{\mu_p C_{ox}(\frac{W}{L})_{3,4}(V_{DD} - V_{cont} - |V_{THP}|)} \leq 0.4(V_{DD} - V_{cont} - |V_{THP}|). \tag{14.66}$$

It follows that

$$V_{cont} \leq V_{DD} - |V_{THP}| - \sqrt{\frac{I_{SS}}{0.4\mu_p C_{ox}(\frac{W}{L})_{3,4}}}. \quad (14.67)$$

If V_{cont} exceeds this level by a large margin, M_3 and M_4 eventually enter saturation. Each stage then requires common-mode feedback to produce the output swings around a well-defined CM level.

The differential pair of Fig. 14.42 suffers from a critical drawback: the output swing of the circuit varies considerably across the tuning range. With complete switching, each stage provides a differential output swing of $2I_{SS}R_{on3,4}$. Thus, a tuning range of, say, two to one translates to a twofold variation in the swing.

In order to minimize the swing variation, the tail current can be adjusted by V_{cont} as well such that, as V_{cont} becomes more positive, I_{SS} decreases. The circuit nonetheless requires a means of maintaining $I_{SS}R_{on3,4}$ relatively constant. To this end, let us consider the circuit in Fig. 14.43(a), where M_5 operates in the deep triode region and amplifier A_1 applies

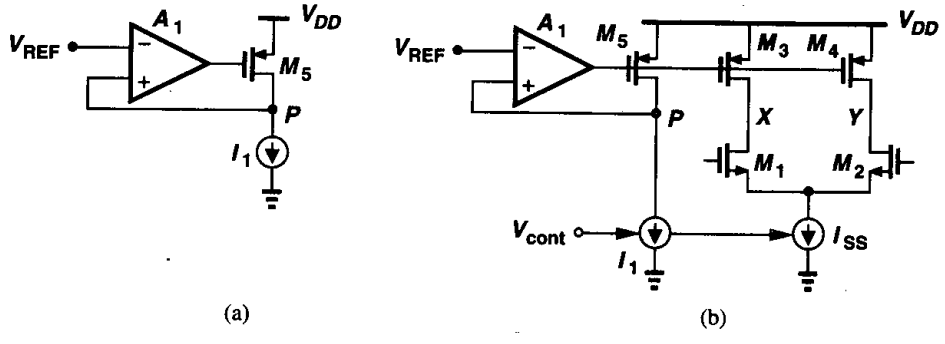


Figure 14.43 (a) Simple feedback circuit defining V_P, (b) replica biasing to define voltage swings in a ring oscillator.

negative feedback to the gate of M_5. If the loop gain is sufficiently large, the differential input voltage of A_1 must be small, giving $V_P \approx V_{REF}$ and $|V_{DS5}| \approx V_{DD} - V_{REF}$. Thus, the feedback ensures a relatively constant drain-source voltage even if I_1 varies. In fact, as I_1, say, decreases, A_1 raises the gate voltage of M_5 such that $R_{on5}I_1 \approx V_{DD} - V_{REF}$.

The topology of Fig. 14.43(a) can serve as a "replica circuit" for the stages of a ring oscillator, thereby defining the oscillation amplitude. Illustrated in Fig. 14.43(b), the idea is to "servo" the on-resistance of M_3 and M_4 to that of M_5 and vary the frequency by adjusting I_1 and I_{SS} simultaneously [2]. If M_3 and M_4 are identical to M_5 and I_{SS} to I_1, then V_X and V_Y vary from V_{DD} to $V_{DD} - V_{REF}$ as M_1 and M_2 steer the tail current to one side or the other. Thus, if process and temperature variations, say, decrease, I_1 and I_{SS}, then A_1 increases the on-resistance of M_3-M_5, forcing V_P and hence V_X and V_Y (when M_1 or M_2 is fully on) equal to V_{REF}.

The bandwidth of the op amp A_1 in Fig. 14.43(b) is of some concern. If a change in V_{cont} takes a long time to change ω_{out}, then the settling speed of a PLL using this VCO degrades significantly (Chapter 15).

Example 14.11

How does the oscillation frequency depend on I_{SS} for a VCO incorporating the stage of Fig. 14.43(b).

Solution

Noting that $R_{on3,4}I_{SS} \approx V_{DD} - V_{REF}$, we have $R_{on3,4} \approx (V_{DD} - V_{REF})/I_{SS}$ and hence

$$f_{osc} \propto \frac{1}{R_{on3,4}C_L} \tag{14.68}$$

$$\propto \frac{I_{SS}}{(V_{DD} - V_{REF})C_L}. \tag{14.69}$$

Thus, the characteristic is relatively linear.

Delay Variation by Positive Feedback To arrive at another tuning technique, recall that a cross-coupled transistor pair such as that of Fig. 14.36 exhibits a negative resistance of $-2/g_m$, a value that can be controlled by the bias current. A negative resistance $-R_N$ placed in parallel with a positive resistance $+R_P$ gives an equivalent value $+R_N R_P/(R_N - R_P)$, which is more positive if $|-R_N| > |+R_P|$. This idea can be applied to each stage of a ring oscillator as illustrated in Fig. 14.44(a). Here, the load of the differential pair consists of

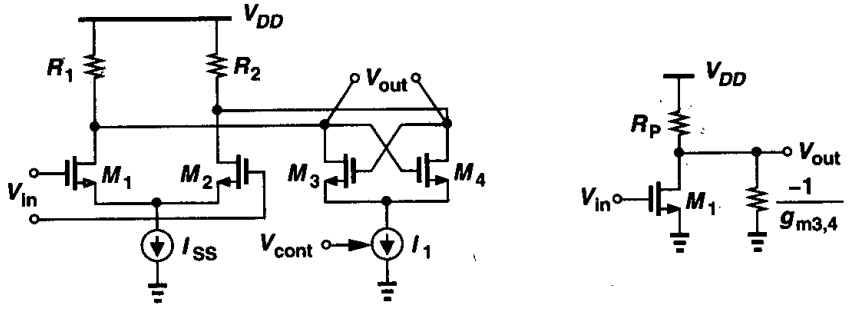


Figure 14.44 (a) Differential stage with variable negative-resistance load, (b) half-circuit equivalent of (a).

resistors R_1 and R_2 ($R_1 = R_2 = R_P$) and the cross-coupled pair M_3-M_4. As I_1 increases, the small-signal differential resistance $-2/g_{m3,4}$ becomes less negative and, from the half circuit of Fig. 14.44(b), the equivalent resistance $R_P||(-1/g_{m3,4}) = R_P/(1 - g_{m3,4}R_P)$ increases, thereby lowering the frequency of oscillation.

An important issue in the circuit of Fig. 14.44(a) is that as I_1 varies, so do the currents steered by M_3 and M_4 to R_1 and R_2. Thus, the output voltage swing is not constant across

the tuning range. To minimize this effect, I_{SS} can be varied in the *opposite* direction such that the total current steered between R_1 and R_2 remains constant. In other words, it is desirable to vary I_1 and I_{SS} *differentially* while their sum is fixed, a characteristic provided by a differential pair. Illustrated in Fig. 14.45, the idea is to employ a differential pair M_5-M_6 to steer I_T to M_1-M_2 or M_3-M_4 so that $I_{SS} + I_1 = I_T$. Since I_T must flow through R_1 and

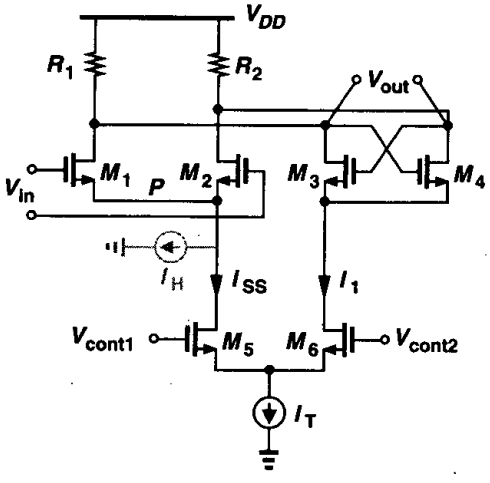


Figure 14.45 Use of a differential pair to steer current between M_1-M_2 and M_3-M_4.

R_2, if M_1-M_4 experience complete switching in each cycle of oscillation, then I_T is steered to R_1 (through M_1 and M_3) in half a period and to R_2 (through M_2 and M_4) in the other half, giving a differential swing of $2R_P I_T$.

In the circuit of Fig. 14.45, V_{cont1} and V_{cont2} can be viewed as differential control lines if they vary by equal and opposite amounts. Such a topology provides higher noise immunity for the control input than if V_{cont} is single-ended. Now, note that as V_{cont1} decreases and V_{cont2} increases, the cross-coupled pair exhibits a greater transconductance, thereby raising the time constant at the output nodes. But what happens if all of I_T is steered by M_6 to M_3 and M_4? Since M_1 and M_2 carry no current, the gain of the stage falls to zero, prohibiting oscillation. To avoid this effect, a small constant current source, I_H, can be connected from node P to ground, thereby ensuring M_1 and M_2 always remain on. With typical values, this ring oscillator provides a two-to-one tuning range and reasonable linearity.

Example 14.12

Calculate the minimum value of I_H in Fig. 14.45 to guarantee a low-frequency gain of 2 when all of I_T is steered to the cross-coupled pair.

Solution

The small-signal voltage gain of the circuit equals $g_{m1,2}R_P/(1 - g_{m3,4}R_P)$. Assuming square-law devices, we have

$$\sqrt{\mu_n C_{ox}(\frac{W}{L})_{1,2} I_H}\,\frac{R_P}{1-\sqrt{\mu_n C_{ox}(\frac{W}{L})_{3,4} I_T R_P}} \geq 2. \tag{14.70}$$

That is,

$$I_H \geq \frac{4\left[1-\sqrt{\mu_n C_{ox}(\frac{W}{L})_{3,4} I_T R_P}\right]^2}{\mu_n C_{ox}(\frac{W}{L})_{1,2} R_P^2}. \tag{14.71}$$

An important drawback of using the differential pair M_5-M_6 in the circuit of Fig. 14.45 is the additional voltage headroom that it consumes. As depicted in Fig. 14.46, for M_5 to remain in saturation, V_P must be sufficiently higher than V_N. When $V_{cont1} = V_{cont2}$,

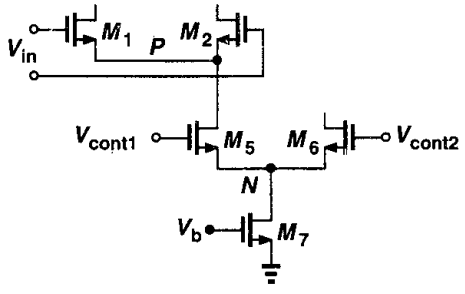


Figure 14.46 Headroom calculation for a current-steering topology.

the minimum allowable drain-source voltage of M_5 is equal to its equilibrium overdrive voltage, implying that, compared to that calculated in Example 14.4, the supply voltage must be higher by this value. Note also that if V_{cont1} or V_{cont2} is allowed to vary above its equilibrium value by more than V_{TH}, then M_5 or M_6 enters the triode region.

The above observation reveals a trade-off between voltage headroom and the *sensitivity* of the VCO. In order to minimize the sensitivity with a given tuning range, the transconductance of M_5-M_6 must be *minimized.* (That is, to steer all of the tail current, the differential pair must require a *large* $V_{cont1} - V_{cont2}$.) However, for a given tail current, $g_m = 2I_D/(V_{GS} - V_{TH})$, indicating a large equilibrium overdrive for M_5-M_6 and a correspondingly higher value for the minimum required supply voltage.

We should mention that the pair M_5-M_6 need not remain in complete saturation. If the drain voltages are low enough to drive these transistors into the triode region, then the equivalent transconductance of the differential pair drops, thus demanding a greater $V_{cont1} - V_{cont2}$ to steer the tail current. This phenomenon in fact translates to a *lower* VCO sensitivity. In practice, careful simulations are required to ensure the VCO characteristic remains relatively linear across the range of interest.[7]

At low supply voltages, it is desirable to avoid the voltage headroom consumed by M_5-M_6 in Fig. 14.45. The issue can be resolved by means of "current folding." Suppose, as illustrated in Fig. 14.47(a), a differential pair drives two current mirrors, generating I_{out1} and I_{out2}. Since $I_1 + I_2 = I_{SS}$, $I_{out1} = KI_1$, and $I_{out2} = KI_2$, we have $I_{out1} + I_{out2} = KI_{SS}$. Thus, as $V_{in1} - V_{in2}$ goes from a very negative value to a very positive value, I_{out1} varies

[7] If both M_5 and M_6 are in the triode region and $V_{cont1} \neq V_{cont2}$, then supply voltage variations affect the current steered between the two transistors, introducing noise in the frequency of oscillation.

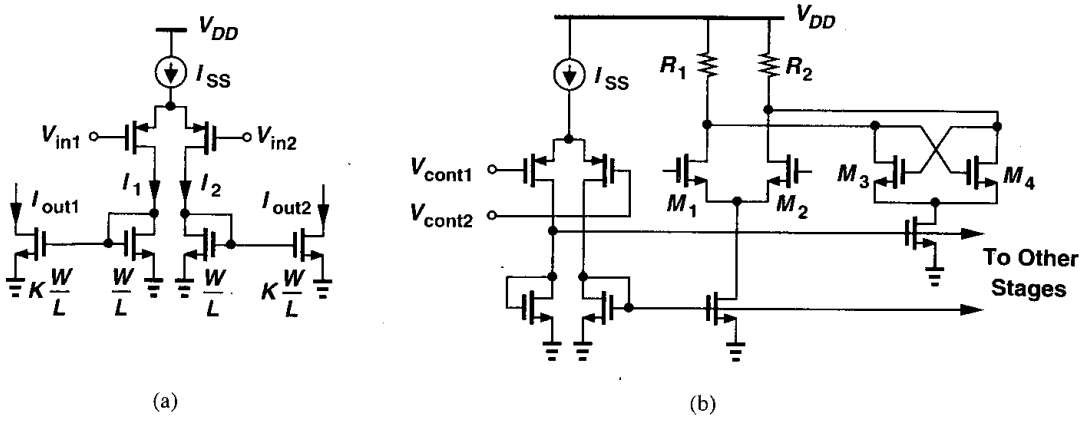
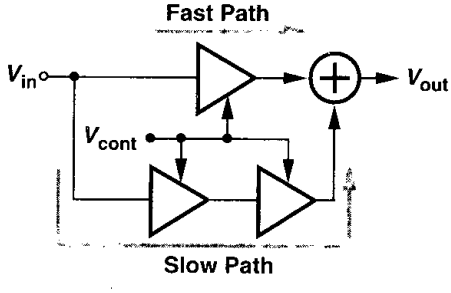


Figure 14.47 (a) Current folding topology, (b) application of current folding to current steering.

from $K I_{SS}$ to zero and I_{out2} from zero to $K I_{SS}$ while their sum remains constant - a behavior similar to that of a differential pair.

We now utilize the topology of Fig. 14.47(a) in the gain stage of Fig. 14.44(a). Shown in Fig. 14.47(b), the resulting circuit operates from a low supply voltage.

Delay Variation by Interpolation Another approach to tuning ring oscillators is based on "interpolation" [3, 4]. As illustrated in Fig. 14.48(a), each stage consists of a fast path and



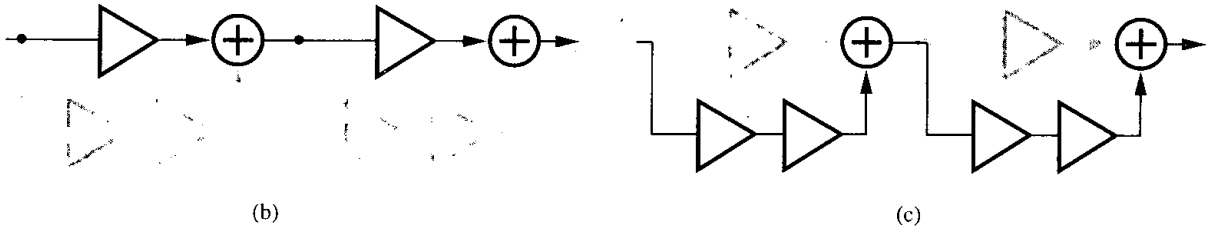


Figure 14.48 (a) Interpolating delay stage, (b) smallest delay, (b) largest delay.

a slow path whose outputs are summed and whose gains are adjusted by V_{cont} in opposite directions. At one extreme of the control voltage, only the fast path is on and the slow path is disabled, yielding the maximum oscillation frequency [Fig. 14.48(b)]. Conversely, at the other extreme, only the slow path is on and the fast path is off, providing the minimum oscillation frequency [Fig. 14.48(c)]. If V_{cont} lies between the two extremes, each path is partially on and the total delay is a weighted sum of their delays.

To better understand the concept of interpolation, let us implement the topology of Fig. 14.48(a) at the transistor level. Each stage can be simply realized as a differential pair whose gain is controlled by its tail current. But how are the two outputs summed? Since the two transistors in a differential pair provide output *currents*, the outputs of the two pairs can be added in the current domain. As depicted in Fig. 14.49(a), simply shorting the outputs of two pairs performs the current addition, e.g., for small signals, $I_{out} = g_{m1,2}V_{in1} + g_{m3,4}V_{in2}$. The overall interpolating stage therefore assumes the configuration shown in Fig. 14.49(b), where V_{cont}^{+} and V_{cont}^{-} denote voltages that vary in opposite directions (so that when one path turns on, the other turns off). The output currents of M_1-M_2 and M_3-M_4 are summed at X and Y and flow through R_1 and R_2, producing V_{out}.

In the circuit of Fig. 14.49(b), the gain of each stage is varied by the tail current to achieve interpolation. But it is desirable to maintain constant voltage swings. We also recognize that the gain of the differential pair M_5-M_6 need not be varied because even if only the gain of M_3-M_4 drops to zero, the slow path is fully disabled. We then surmise that if the tail currents of M_1-M_2 and M_3-M_4 vary in opposite directions such that their sum remains constant, we achieve both interpolation between the two paths and constant output swings.

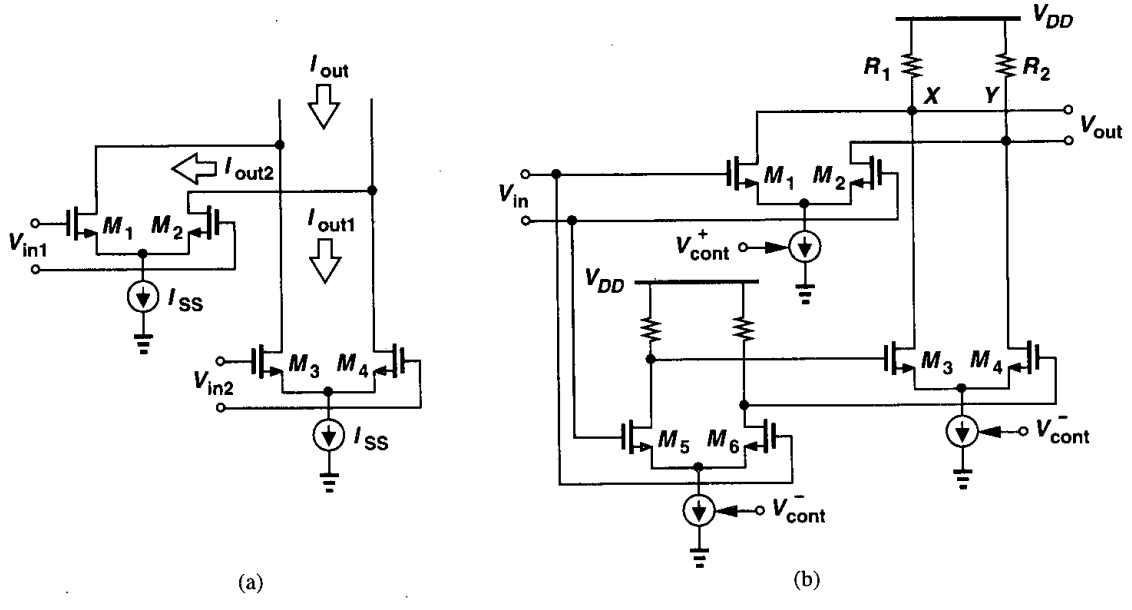


Figure 14.49 (a) Addition of currents of two differential pairs, (b) interpolating delay stage.

Illustrated in Fig. 14.50, the resulting circuit employs the differential pair M_7-M_8 to steer I_{SS} between M_1-M_2 and M_3-M_4. If V_{cont} is very negative, M_8 is off and only the fast path

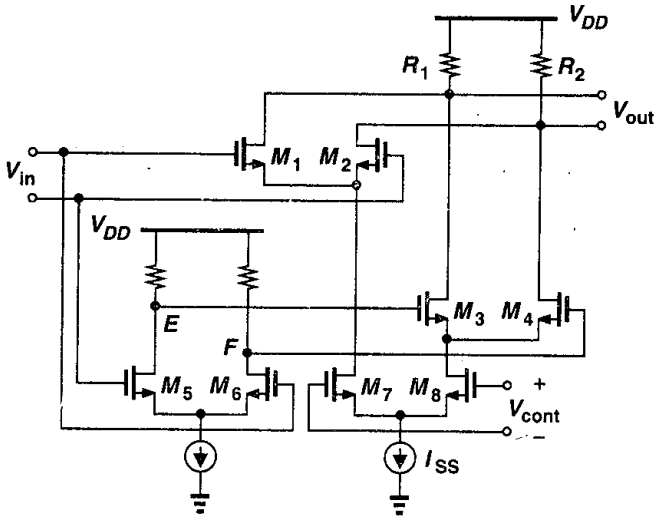


Figure 14.50 Interpolating delay stage with current steering.

amplifies the input. Conversely, if V_{cont} is very positive, M_7 is off and only the slow path is enabled. Since the slow path in this case employs one more stage than the fast path, the VCO achieves a tuning range of roughly two to one. For operation with low supply voltages, the control pair M_7-M_8 can be replaced by the current-folding topology of Fig. 14.47(a).

Example 14.13

Combine the tuning techniques of Figs. 14.45 and 14.50 to achieve a wider tuning range.

Solution

We begin with the interpolating stage of Fig. 14.50 and add a cross-coupled pair to the output nodes [Fig. 14.51(a)]. However, in order to obtain constant voltage swings, the total current through the load resistors must remain constant. This is accomplished by replacing the control differential pair with the current-folding circuit of Fig. 14.47(a). Depicted in Fig. 14.51(b), the resulting configuration steers the current to M_1-M_2 to speed up the circuit and to M_3-M_4 and M_{10}-M_{11} to slow down the circuit. The tail current source dimensions are chosen such that $I_{SS1} = I_{SS2} + I_{SS3}$.

Wide-Range Tuning Except for the circuit of Fig. 14.43(b), the ring oscillator tuning techniques presented thus far achieve a tuning range of typically no more than three to one. In applications where the frequency must be varied by orders of magntitude, the topology shown in Fig. 14.52 can be used. Driven by the input, the additional PMOS transistors M_5 and M_6 pull each output node to V_{DD}, creating a relatively constant output swing even with large variations in I_{SS}. The oscillation frequency of a ring incorporating this stage can be varied by more than four orders of magnitude with less than a twofold variation in the amplitude.

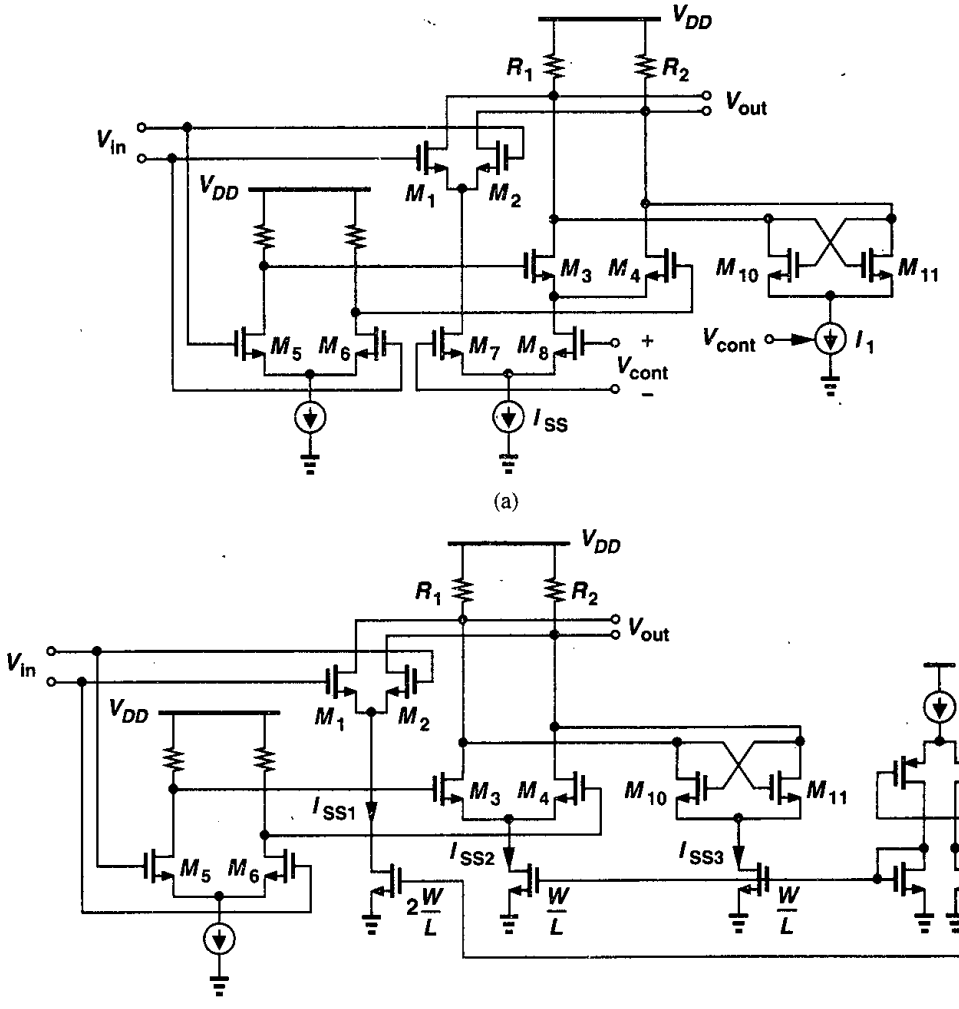


Figure 14.51

14.4.2 Tuning in LC Oscillators

The oscillation frequency of LC topologies is equal to $f_{osc} = 1/(2\pi\sqrt{LC})$, suggesting that only the inductor and capacitor values can be varied to tune the frequency and other parameters such as bias currents and transistor transconductances affect f_{osc} negligibly. Since it is difficult to vary the value of monolithic inductors, we simply change the tank capacitance to tune the oscillator. Voltage-dependent capacitors are called "varactors."[8]

[8]The term "varicap" is also used.

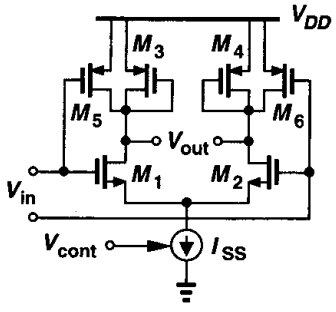


Figure 14.52 Differential stage with wide tuning range.

A reverse-biased *pn* junction can serve as a varactor. The voltage dependence is expressed as

$$C_{var} = \frac{C_0}{(1 + \frac{V_R}{\phi_B})^m}, \tag{14.72}$$

where C_0 is the zero-bias value, V_R the reverse-bias voltage, ϕ_B the built-in potential of the junction, and m a value typically between 0.3 and 0.4.[9] Equation (14.72) reveals an important drawback of LC oscillators: at low supply voltages V_R has a very limited range, yielding a small range for C_{var} and hence for f_{osc}. We also note that to maximize the tuning range, constant capacitances in the tank must be *minimized.*

Example 14.14

Suppose in Eq. (14.72), $\phi_B = 0.7$ V, $m = 0.35$, and V_R can vary from zero to 2 V. How much tuning range can be achieved?

Solution

For $V_R = 0$, $C_j = C_0$ and $f_{osc,min} = 1/(2\pi\sqrt{LC_0})$. For $V_R = 2$ V, $C_j \approx 0.62C_0$ and $f_{osc,max} = 1/(2\pi\sqrt{L \times 0.62C_0}) \approx 1.27 f_{osc,min}$. Thus, the tuning range is approximately equal to 27%. As explained later, the parasitic capacitances of the inductor and the transistor(s) further limit this range because they cannot be varied by the control voltage.

Let us now add varactor diodes to a cross-coupled LC oscillator (Fig. 14.53). To avoid forward-biasing D_1 and D_2 significantly, V_{cont} must not exceed V_X or V_Y by more than a few hundred millivolts. Thus, if the peak amplitude at each node is A, then $0 < V_{cont} < V_{DD} - A + 300$ mV, where it is assumed a forward bias of 300 mV creates negligible current. Interestingly, the circuit suffers from a trade-off between the output swing and the tuning range. This effect appears in most LC oscillators.

Note that, since the swings at X and Y are typically large (e.g., 1 V_{pp} at each node), the capacitance of D_1 and D_2 *varies* with time. Nonetheless, the "average" value of the capacitance is still a function of V_{cont}, providing the tuning range.

[9]Note that $m = 0.5$ for an abrupt junction, but *pn* junctions in CMOS technology are not abrupt.

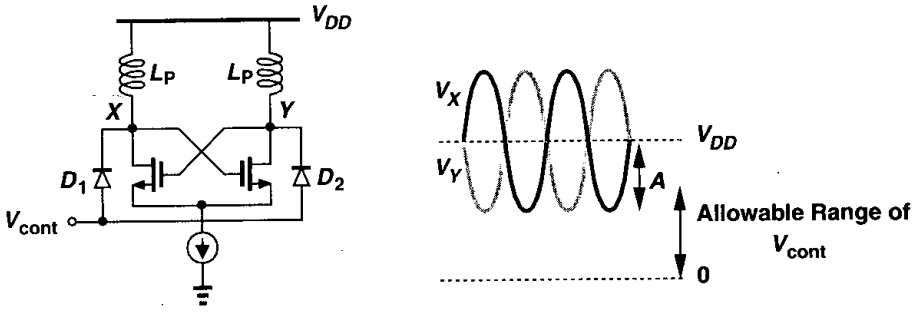


Figure 14.53 LC oscillator using varactor diodes.

How are varactor diodes realized in CMOS technology? Illustrated in Fig. 14.54 are two types of *pn* junctions. In Fig. 14.54(a), the anode is inevitably grounded whereas in Fig.

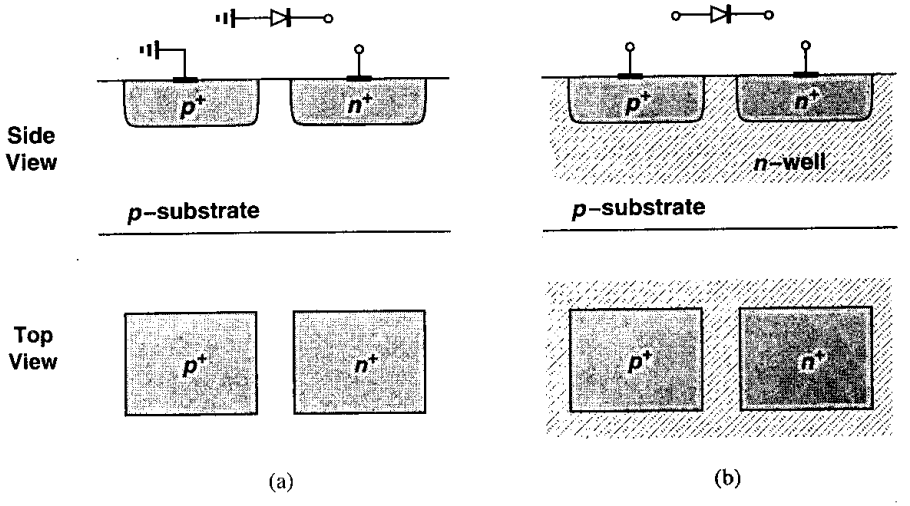


Figure 14.54 Diodes realized in CMOS technology.

14.54(b), both terminals are floating. For the circuit of Fig. 14.53, only the floating diode can be used. To increase the capacitance of the junction, the p^+ and n^+ areas (and hence the n-well) are enlarged.

Upon closer examination, the structure of Fig. 14.54(b) suffers from a number of drawbacks. First, the n-well material has a high resistivity, creating a resistance in series with the reverse-biased diode and lowering the quality factor of the capacitance. Second, the n-well displays substantial capacitance to the substrate, contributing a constant capacitance to the tank and limiting the tuning range. The diode is therefore represented as shown in Fig. 14.55, where C_n represents the (voltage-dependent) capacitance between the n-well and the substrate.[10]

[10] In circuit simulations, C_n is replaced by a diode having proper junction capacitance.

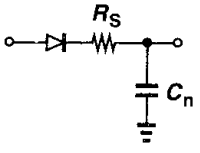


Figure 14.55 Circuit model of the varactor shown in Fig. 14.54(b).

In order to decrease the series resistance of the structure shown in Fig. 14.54(b), the p^+ region can be surrounded by an n^+ ring so that the displacement current flowing through the junction capacitance sees a low resistance in all four directions [Fig. 14.56(a)]. Since a single minimum-size p^+ area has a small capacitance, many of these units can be placed in parallel [Fig. 14.56(b)]. The n-well, however, must accommodate the entire set, exhibiting a large capacitance to the substrate.

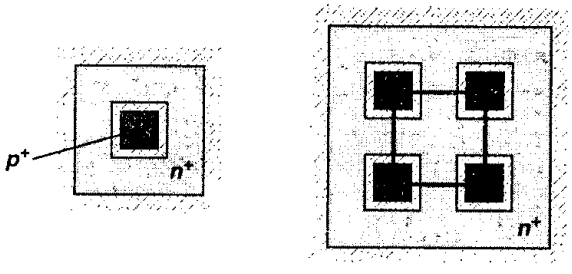


Figure 14.56 (a) Reduction of series resistance by surrounding the p^+ region by an n^+ ring, (b) several diodes in parallel.

It is instructive at this point to examine the unwanted capacitances in the circuit of Fig. 14.53, i.e., the components that are not varied by V_{cont}. We identify three such capacitances: (1) the capacitance between the n-well and the substrate associated with D_1 and D_2; (2) the capacitances contributed by the transistors to each node, i.e., C_{GD}, $2C_{GD}$ (the factor of 2 arising from Miller effect[11]), and C_{DB}; (3) the parasitic capacitance of the inductor itself. Monolithic inductors are typically implemented as metal spiral structures (Fig. 14.57) having relatively large dimensions ($S \approx$ 100-200 μm). Their capacitance to the substrate is therefore quite large.

In Fig. 14.53, it is desirable to connect the anode of the diodes to nodes X and Y, thereby eliminating the parasitic n-well capacitances from the tank. Shown in Fig. 14.58 is a topology allowing such a modification. Here, the cross-coupled pair incorporates PMOS devices, providing swings around the ground potential. However, owing to their lower mobility, the PMOS transistors must be wider than their NMOS counterparts so as to exhibit the same transconductance. This increases the second component mentioned above.

[11] If the gate and drain voltages vary by equal and opposite amounts, the Miller multiplication factor is equal to 2 regardless of the small-signal gain.

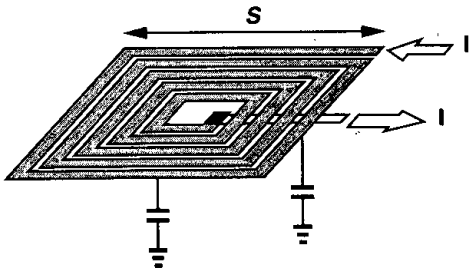


Figure 14.57 Spiral inductor structure.

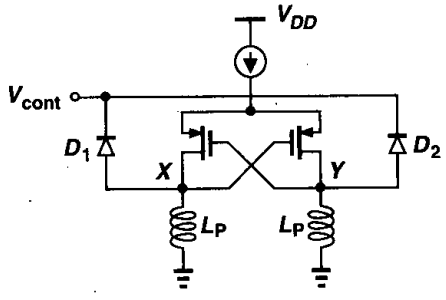


Figure 14.58 Negative-G_m oscillator using PMOS devices to eliminate n-well capacitance from the tanks.

The design of low-noise CMOS LC oscillators with acceptable tuning range is still a topic of active research. Issues such as phase noise and inductor and varactor design continue to intrigue researchers. MOS varactors have also been investigated as an alternative to *pn* junctions.

14.5 Mathematical Model of VCOs

The definition of the voltage-controlled oscillator given by Eq. (14.55) specifies the relationship between the control voltage and the output frequency. The dependence is "memoryless" because a change in V_{cont} immediately results in a change in ω_{out}. But how is the output signal of the VCO expressed as a function of time? To answer this question, we must review the concepts of phase and frequency.

Consider the waveform $V_0(t) = V_m \sin \omega_0 t$. The argument of the sinusoid is called the "total phase" of the signal. In this example, the phase varies linearly with time, exhibiting a slope equal to ω_0. Note that, as depicted in Fig. 14.59, every time $\omega_0 t$ crosses an integer multiple of π, $V_0(t)$ crosses zero.

Now consider two waveforms $V_1(t) = V_m \sin[\phi_1(t)]$ and $V_2(t) = V_m \sin[\phi_2(t)]$, where $\phi_1(t) = \omega_1 t$, $\phi_2(t) = \omega_2 t$, and $\omega_1 < \omega_2$. As illustrated in Fig. 14.60, $\phi_2(t)$ crosses integer multiples of π faster than $\phi_1(t)$ does, yielding faster variations in $V_2(t)$. We say $V_2(t)$ accumulates phase faster.

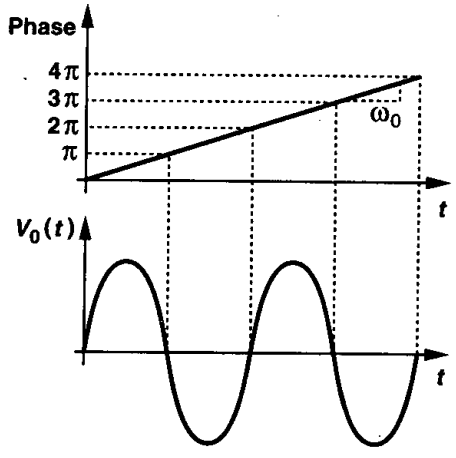


Figure 14.59 Illustration of phase of a signal.

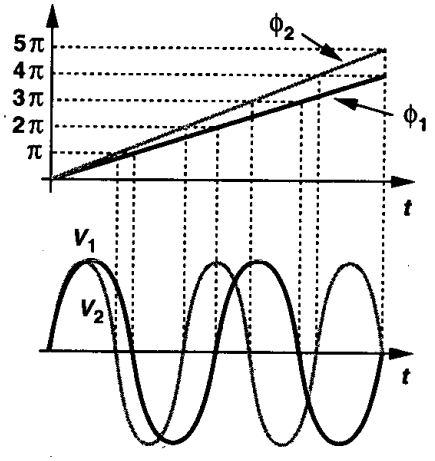


Figure 14.60 Variation of phase for two signals.

The above study reveals that the faster the phase of a waveform varies, the higher the frequency of the waveform, suggesting that the frequency[12] can be defined as the derivative of the phase with respect to time:

$$\omega = \frac{d\phi}{dt}. \tag{14.73}$$

Example 14.15

Figure 14.61(a) shows the phase of a sinusoidal waveform with constant amplitude as a function of time. Plot the waveform in the time domain.

[12]The quantity $\omega = 2\pi f$ is called the "radian frequency" (and expressed in rad/s) to distinguish it from f (expressed in Hz). In this book, we call both the frequency, but use ω more often to avoid the factor 2π.

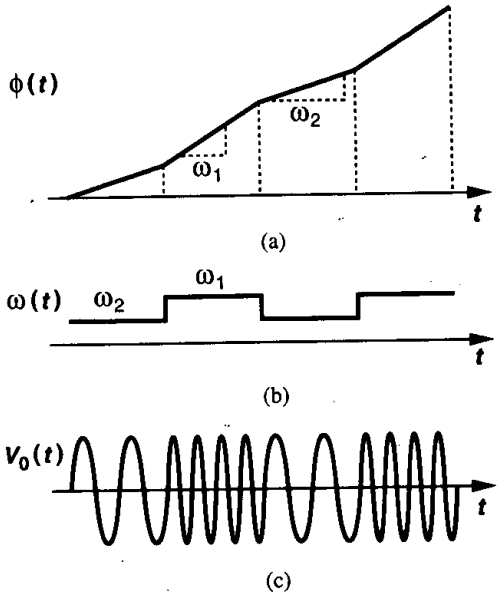


Figure 14.61

Solution

Taking the time derivative of $\phi(t)$, we obtain the behavior illustrated in Fig. 14.61(b). The frequency therefore periodically toggles between ω_1 and ω_2, yielding the waveform shown in Fig. 14.61(c). (This is a simple example of binary frequency modulation, called "frequency shift keying" and utilized in wireless pagers and many other communication systems.)

Equation (14.73) indicates that, if the frequency of a waveform is known as a function of time, then the phase can be computed as

$$\phi = \int \omega dt + \phi_0. \tag{14.74}$$

In particular, since for a VCO, $\omega_{out} = \omega_0 + K_{VCO} V_{cont}$, we have

$$V_{out}(t) = V_m \cos(\int \omega_{out} dt + \phi_0) \tag{14.75}$$

$$= V_m \cos(\omega_0 t + K_{VCO} \int V_{cont} dt + \phi_0). \tag{14.76}$$

Equation (14.76) proves essential in the analysis of VCOs and PLLs.[13] The initial phase ϕ_0 is usually unimportant and is assumed zero hereafter.

[13]Note that K_{VCO} cannot be brought out of the integral if the characteristic is nonlinear.

Example 14.16

The control line of a VCO senses a rectangular signal toggling between V_1 and V_2 at a period T_m. Plot the frequency, phase, and output waveform as a function of time.

Solution

Since $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$, the output frequency toggles between $\omega_1 = \omega_0 + K_{VCO}V_1$ and $\omega_2 = \omega_0 + K_{VCO}V_2$ (Fig. 14.62). The phase is equal to the time integral of this result, rising linearly

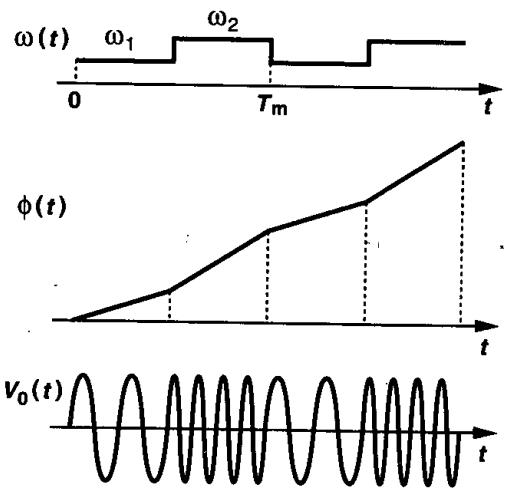


Figure 14.62

with time at a slope of ω_1 for half the input period and ω_2 for the other half. The output waveform of the VCO is similar to that shown in Fig. 14.61. Thus, a VCO can operate as a frequency modulator.

As explained in Chapter 15, if a VCO is placed in a phase-locked loop, then only the second term of the total phase in Eq. (14.76) is of interest. This term, $K_{VCO}\int V_{cont}dt$, is called the "excess phase," ϕ_{ex}. In fact, in the analysis of PLLs, we view the VCO as a system whose input and output are the control voltage and the excess phase, respectively:

$$\phi_{ex} = K_{VCO}\int V_{cont}dt. \tag{14.77}$$

That is, the VCO operates as an *ideal* integrator, providing a transfer function:

$$\frac{\Phi_{ex}}{V_{cont}}(s) = \frac{K_{VCO}}{s}. \tag{14.78}$$

Example 14.17

A VCO senses a small sinusoidal control voltage $V_{cont} = V_m \cos \omega_m t$. Determine the output waveform and its spectrum.

Solution

The output is expressed as

$$V_{out}(t) = V_0 \cos(\omega_0 t + K_{VCO} \int V_{cont} dt) \tag{14.79}$$

$$= V_0 \cos(\omega_0 t + K_{VCO} \frac{V_m}{\omega_m} \sin \omega_m t) \tag{14.80}$$

$$= V_0 \cos \omega_0 t \cos(K_{VCO} \frac{V_m}{\omega_m} \sin \omega_m t) \tag{14.81}$$

$$-V_0 \sin \omega_0 t \sin(K_{VCO} \frac{V_m}{\omega_m} \sin \omega_m t).$$

If V_m is small enough that $K_{VCO} V_m / \omega_m \ll 1$ rad, then

$$V_{out}(t) \approx V_0 \cos \omega_0 t - V_0 (\sin \omega_0 t)(K_{VCO} \frac{V_m}{\omega_m} \sin \omega_m t) \tag{14.82}$$

$$= V_0 \cos \omega_0 t - \frac{K_{VCO} V_m V_0}{2\omega_m} [\cos(\omega_0 - \omega_m)t - \cos(\omega_0 + \omega_m)t]. \tag{14.83}$$

The output therefore consists of three sinusoids having frequencies of ω_0, $\omega_0 - \omega_m$, and $\omega_0 + \omega_m$. The spectrum is shown in Fig. 14.63. The components at $\omega_0 \pm \omega_m$ are called "sidebands."

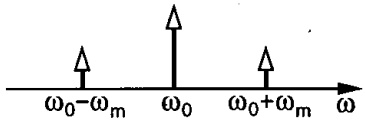


Figure 14.63

The above example reveals that variation of the control voltage with time may create unwanted components at the output. Indeed, when a VCO operates in the steady state, the control voltage must experience very little variation.[14] This issue is studied in Chapter 15.

A common mistake in expressing the phase of signals arises from the familiar form $V_m \cos \omega_0 t$. Here, the phase is equal to the product of frequency and time, creating the impression that such equality holds in all conditions. We may even deduce that, since the output frequency of a VCO is given by $\omega_0 + K_{VCO} V_{cont}$, the output waveform can be written as $V_m \cos[(\omega_0 + K_{VCO} V_{cont})t]$. To understand why this is incorrect, let us compute the frequency as the derivative of the phase:

$$\omega = \frac{d}{dt}[(\omega_0 + K_{VCO} V_{cont})t] \tag{14.84}$$

$$= K_{VCO} \frac{dV_{cont}}{dt} t + \omega_0 + K_{VCO} V_{cont}. \tag{14.85}$$

[14] Except when the VCO senses a signal to perform frequency modulation.

The first term in this expression is redundant, vanishing only if $dV_{cont}/dt = 0$. Thus, in the general case, the phase cannot be written as the product of time and frequency.

Our study of VCOs in this section has assumed sinusoidal output waveforms. In practice, depending on the type and speed of the oscillator, the output may contain significant harmonics, even approaching a rectangular waveform. How should Eq. (14.76) be modified in this case? We expect that $V_{out}(t)$ can be expressed as a Fourier series:

$$V_{out}(t) = V_1 \cos(\omega_0 t + \phi_1) + V_2 \cos(2\omega_0 t + \phi_2) + \cdots. \tag{14.86}$$

We also note that if the (fundamental) frequency of a rectagular waveform is changed by Δf, the frequency of its second harmonic must change by $2\Delta f$, etc. Thus, if V_{cont} varies by ΔV, then the frequency of the first harmonic varies by $K_{VCO}\Delta V$, the frequency of the second harmonic by $2K_{VCO}\Delta V$, etc. That is,

$$V_{out}(t) = V_1 \cos(\omega_0 t + K_{VCO} \int V_{cont} dt + \theta_1) + V_2 \cos(2\omega_0 t + 2K_{VCO} \int V_{cont} dt + \theta_2) + \cdots, \tag{14.87}$$

where $\theta_1, \theta_2, \cdots$ are constant phases necessary for the representation of each harmonic in the Fourier series expansion.

Equation (14.87) suggests that the harmonics of an oscillator output can be readily taken into account. For this reason, we often limit our calculations to the first harmonic even though we may draw the waveforms in rectangular shape rather than sinusoidal shape.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume $V_{DD} = 3$ V where necessary. Also, assume all transistors are in saturation.

14.1. For the circuit of Fig. 14.6, determine the open-loop tranfer function and calculate the phase margin. Assume $g_{m1} = g_{m2} = g_m$ and neglect other capacitances.

14.2. In the circuit of Fig. 14.8, assume $g_{m1} = g_{m2} = g_{m3} = (200\ \Omega)^{-1}$.
(a) What is the minimum value of R_D that ensures oscillation?
(b) Determine the value of C_L for an oscillation frequency of 1 GHz and a total low-frequency loop gain of 16.

14.3. For the circuit of Fig. 14.12, determine the minimum value of I_{SS} that guarantees oscillation. (Hint: if the circuit is at the edge of oscillation, the swings are quite small.)

14.4. Prove that the small-signal resistance of the composite load in Fig. 14.18(c) is roughly equal to $1/g_{m3}$.

14.5. Including only the gate-source capacitance of M_3 in Fig. 14.18(c), explain under what condition the impedance of the composite load (seen at the drain of M_3) becomes inductive.

14.6. If each inductor in Fig. 14.25 exhibits a series resistance of R_S, how low must R_S be to ensure the low-frequency loop gain is less than unity? (This condition is necessary to avoid latchup.)

14.7. Explain why the V_X and V_Y waveforms in Fig. 14.28 are closer to sinusoids (i.e., they contain smaller harmonics) than the I_{D1} and I_{D2} waveforms.

14.8. Determine the minimum value of I_{SS} in Fig. 14.27(c) that guarantees oscillation. Estimate the maximum value of I_{SS} that guarantees M_1 and M_2 do not enter the triode region.

14.9. Repeat Example 14.7 by applying a current stimulus to the drain of M_1.

14.10. Prove that if a capacitor C_P is placed in parallel with L_P in Fig. 14.31(a), then Eq. (14.47) results.

14.11. The Colpitts oscillator of Fig. 14.31(a) was analyzed and its oscillation conditions were derived by applying a current stimulus to the source. Repeat the analysis by applying a voltage stimulus to the gate of M_1.

14.12. Repeat the analysis of the Colpitts oscillator for the topologies in Figs. 14.38(a) and (c). Determine the oscillation condition and the frequency of oscillation.

14.13. The stage of Fig. 14.45 is designed with $I_T = 1$ mA and $(W/L)_{1,2} = 50/0.5$. Assume $I_H \ll I_1$.

(a) Determine the minimum value of $R_1 = R_2 = R$ to ensure oscillation in a three-stage ring.

(b) Determine $(W/L)_{3,4}$ such that $g_{m3,4}R = 0.5$ when each of M_3 and M_4 carries $I_T/2$.

(c) Calculate the minimum value of I_H to guarantee oscillation.

(d) If the common-mode level of V_{cont1} and V_{cont2} is 1.5 V, calculate $(W/L)_{5,6}$ such that I_T sustains 0.5 V when $V_{cont1} = V_{cont2}$.

14.14. Repeat Example 14.14 if each inductor in the circuit contributes a constant capacitance equal to C_1.

14.15. The VCO of Fig. 14.53 is designed for operation at 1 GHz.

(a) If $L_P = 5$ nH and the total (fixed) parasitic capacitance seen at X (and Y) to ground is 500 fF, determine the maximum capacitance that D_1 and D_2 can add to the circuit.

(b) If the tail current is equal to 1 mA and the Q of each inductor at 1 GHz is equal to 4, estimate the output voltage swing.

References

1. N. M. Nguyen and R. G. Meyer, "Start-up and Frequency Stability in High-Frequency Oscillators," *IEEE Journal of Solid-State Circuits,* vol. 27, pp. 810–820, May 1992.
2. I. A. Young, J. K. Greason, and K. L. Wong, "A PLL Clock Generator with 5 to 110 MHz of Lock Range for Microprocessors," *IEEE Journal of Solid-State Circuits,* vol. SC-27, pp.1599–1607, Nov. 1992.
3. B. Lai and R. C. Walker, "A Monolithic 622 Mb/sec Clock Extraction and Data Retiming Circuit," *ISSCC Dig. Tech. Papers,* pp. 144–145, Feb. 1991.
4. S. K. Enam and A. A. Abidi, "NMOS ICs for Clock and Data Regeneration in Gigabit-per-Second Optical-Fiber Receivers," *IEEE Journal of Solid-State Circuits,* vol. SC-27, pp. 1763–1774, Dec. 1992.

Chapter 15

Phase-Locked Loops

The concept of phase locking was invented in the 1930s and swiftly found wide usage in electronics and communication. While the basic phase-locked loop has remained nearly the same since then, its implementation in different technologies and for different applications continues to challenge designers. A PLL serving the task of clock generation in a microprocessor appears quite similar to a frequency synthesizer used in a cellphone, but the actual circuits are designed quite differently.

This chapter deals with the analysis and design of PLLs with particular attention to implementations in VLSI technologies. A thorough study of PLLs would require an entire book by itself, but our objective here is to lay the foundation for more advanced work. Beginning with a simple PLL architecture, we study the phenomenon of phase locking and analyze the behavior of PLLs in the time and frequency domains. We then address the problem of lock acquisition and describe charge-pump PLLs (CPPLLs) and their nonidealities. Finally, we examine jitter in PLLs, study delay-locked loops (DLLs), and present a number of PLL applications.

15.1 Simple PLL

A PLL is a feedback system that compares the output phase with the input phase. The comparison is performed by a "phase comparator" or "phase detector" (PD). It is therefore beneficial to define the PD rigorously.

15.1.1 Phase Detector

A phase detector is a circuit whose average output, $\overline{V_{out}}$, is linearly proportional to the phase difference, $\Delta\phi$, between its two inputs (Fig. 15.1). In the ideal case, the relationship between $\overline{V_{out}}$ and $\Delta\phi$ is linear, crossing the origin for $\Delta\phi = 0$. Called the "gain" of the PD, the slope of the line, K_{PD}, is expressed in V/rad.

A familiar example of phase detector is the exclusive OR (XOR) gate. As shown in Fig. 15.2, as the phase difference between the inputs varies, so does the width of the output pulses, thereby providing a dc level proportional to $\Delta\phi$. While the XOR circuit produces

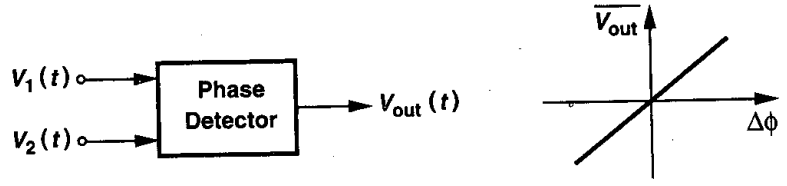


Figure 15.1 Definition of phase detector.

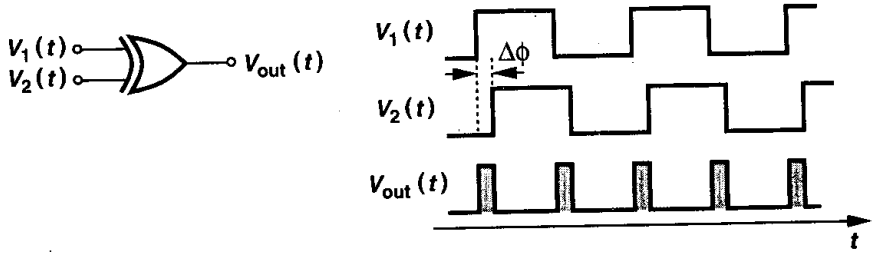


Figure 15.2 Exclusive OR gate as phase detector.

error pulses on both rising and falling edges, other types of PD may respond only to positive or negative transitions.

Example 15.1

If the output swing of the XOR in Fig. 15.2 is V_0 volts, what is the gain of the circuit as a phase detector? Plot the input-output characteristic of the PD.

Solution

If the phase difference increases from zero to $\Delta\phi$ radians, the area under each pulse increases by $V_0 \cdot \Delta\phi$. Since each period contains *two* pulses, the average value rises by $2[V_0 \cdot \Delta\phi/(2\pi)]$, yielding a gain of V_0/π. Note that the gain is independent of the input frequency.

To construct the input-output characteristic, we examine the circuit's response to various input phase differences. As illustrated in Fig. 15.3, the average output voltage rises to $[V_0/\pi] \times \pi/2 = V_0/2$ for $\Delta\phi = \pi/2$ and V_0 for $\Delta\phi = \pi$. For $\Delta\phi > \pi$, the average begins to *drop*, falling to $V_0/2$ for $\Delta\phi = 3\pi/2$ and zero for $\Delta\phi = 2\pi$. The characteristic is therefore periodic, exhibiting both negative and positive gains.

The operation of phase detectors is similar to that of differential amplifiers in that both sense the *difference* between the two inputs, generating a proportional output.

15.1.2 Basic PLL Topology

To arrive at the concept of phase locking, let us consider the problem of aligning the output phase of a VCO with the phase of a reference clock. As illustrated in Fig. 15.4(a), the rising

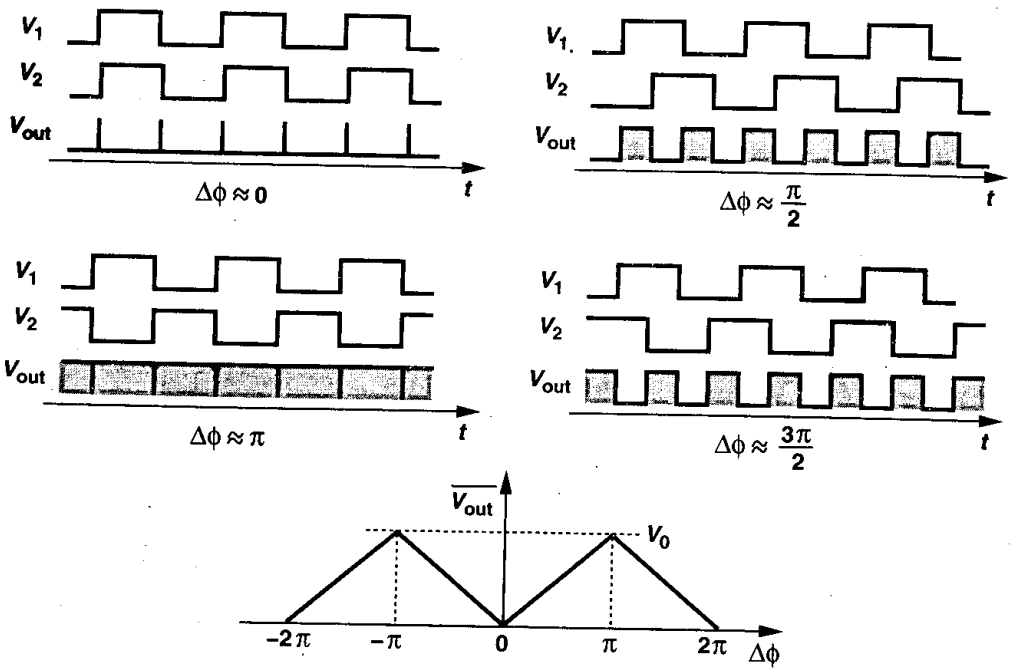


Figure 15.3

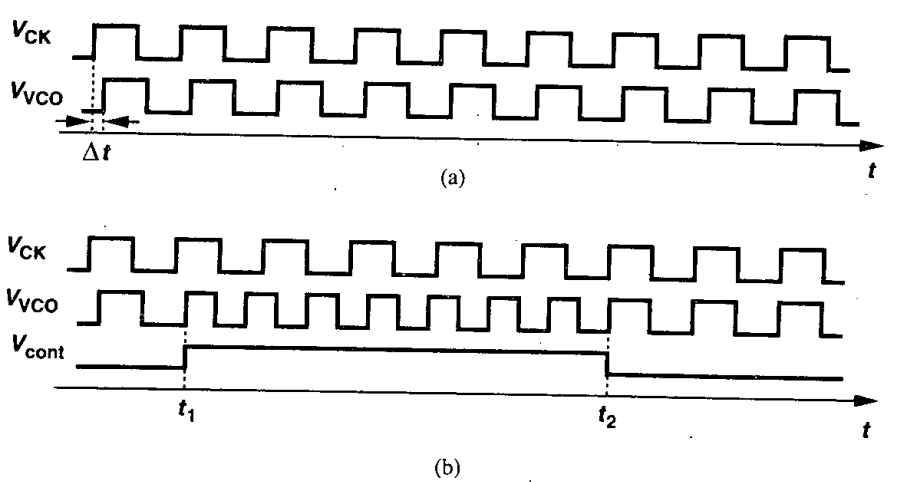


Figure 15.4 (a) Two waveforms with a skew, (b) change of VCO frequency to eliminate the skew.

edges of V_{VCO} are "skewed" by Δt seconds with respect to V_{CK}, and we wish to eliminate this error. Assuming that the VCO has a single control input, V_{cont}, we note that to vary the phase, we *must* vary the frequency and allow the integration $\phi = \int (\omega_0 + K_{VCO} V_{cont})\, dt$ to take place. For example, suppose as shown in Fig. 15.4(b), the VCO frequency is stepped to a higher value at $t = t_1$. The circuit then accumulates phase faster, gradually decreasing the phase error. At $t = t_2$, the phase error drops to zero and, if V_{cont} returns to its original value, V_{VCO} and V_{CK} remain aligned. Interestingly, the alignment can be accomplished by stepping the VCO frequency to a *lower* value for a certain time interval as well (Problem 15.2). Thus, phase alignment can be achieved only by a (temporary) frequency change.

The foregoing experiment suggests that the output phase of a VCO can be aligned with the phase of a reference if (1) the frequency of the VCO is changed momentarily, (2) a means of comparing the two phases, i.e., a phase detector, is used to determine when the VCO and reference signals are aligned. The task of aligning the output phase of the VCO with the phase of the reference is called "phase locking."

From the above observations, we surmise that a PLL simply consists of a PD and a VCO in a feedback loop [Fig. 15.5(a)]. The PD compares the phases of V_{out} and V_{in}, generating an error that varies the VCO frequency until the phases are aligned, i.e., the loop is locked.

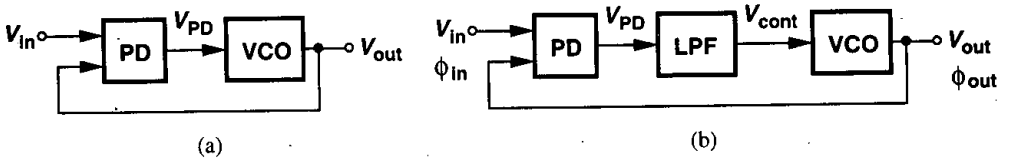


Figure 15.5 (a) Feedback loop comparing input and output phases, (b) simple PLL.

This topology, however, must be modified because (1) as exemplified by the waveforms of Fig. 15.2, the PD output, V_{PD}, consists of a dc component (desirable) and high-frequency components (undesirable), and (2) as mentioned in Chapter 14, the control voltage of the oscillator must remain quiet in the steady state, i.e., the PD output must be filtered. We therefore interpose a low-pass filter (LPF) between the PD and the VCO [Fig. 15.5(b)], suppressing the high-frequency components of the PD output and presenting the dc level to the oscillator. This forms the basic PLL topology. For now, we assume the LPF has a gain of unity at low frequencies (e.g., as in a first-order RC section).

It is important to bear in mind that the feedback loop of Fig. 15.5(b) compares the *phases* of the input and output. Unlike the feedback topologies studied in the previous chapters, PLLs typically require no knowledge of voltages or currents in their feedback operation. If the loop gain is large enough, the difference between the input phase, ϕ_{in}, and the output phase, ϕ_{out}, falls to a small value in the steady state, providing phase alignment.

For subsequent analyses of PLLs, we must define the phase lock condition carefully. If the loop of Fig. 15.5(b) is locked, we postulate that $\phi_{out} - \phi_{in}$ is constant and preferably small. We therefore define the loop to be locked if $\phi_{out} - \phi_{in}$ does not change with time.

An important corollary of this definition is that

$$\frac{d\phi_{out}}{dt} - \frac{d\phi_{in}}{dt} = 0 \tag{15.1}$$

and hence

$$\omega_{out} = \omega_{in}. \tag{15.2}$$

This is a unique property of PLLs and will be revisited more closely later.

In summary, when locked, a PLL produces an output that has a small phase error with respect to the input but exactly the same frequency. The reader may then wonder why a PLL is used at all. A short piece of wire would seem to perform the task even better! We answer this question in Section 15.5.

Example 15.2

Implement a simple PLL in CMOS technology.

Solution

Figure 15.6 illustrates an implementation utilizing an XOR gate as the phase detector. The VCO is

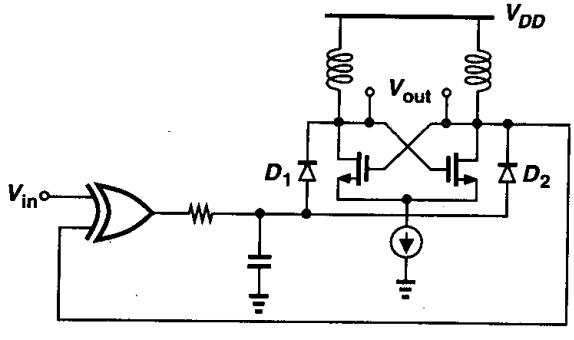


Figure 15.6

configured as a negative-G_m LC oscillator whose frequency is tuned by varactor diodes.

PLL Waveforms in Locked Condition In order to familiarize ourselves with the behavior of PLLs, we begin with the simplest case: the circuit is locked and we wish to examine the waveforms at each point around the loop. As illustrated in Fig. 15.7(a), V_{in} and V_{out} exhibit a small phase difference but equal frequencies. The PD therefore generates pulses as wide as the skew between the input and the output[1] and the low-pass filter extracts the dc component of V_{PD}, applying the result to the VCO. We assume the LPF has a gain of unity at low frequencies. The small pulses in V_{LPF} are called "ripple."

[1] In this example, the PD produces pulses only on the rising transitions.

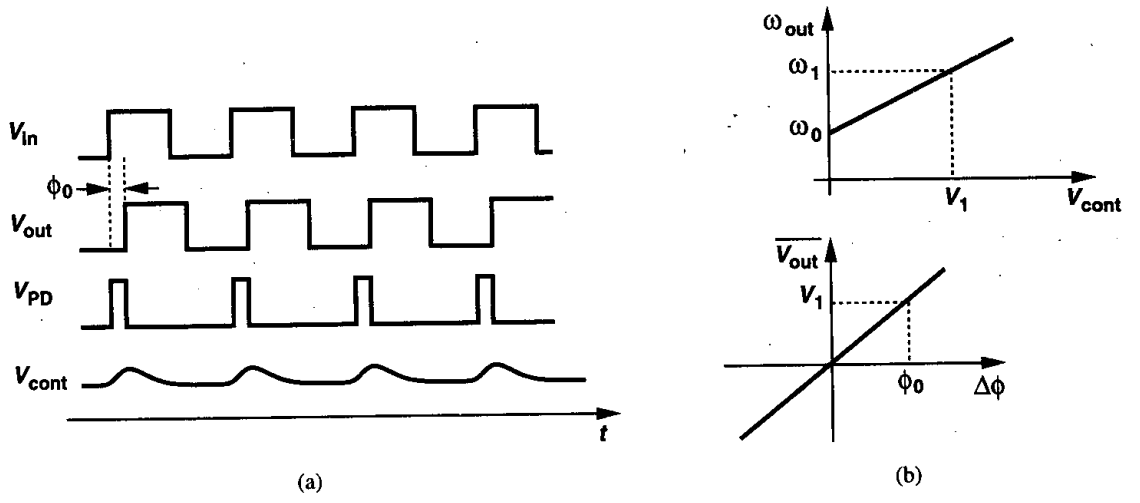


Figure 15.7 (a) Waveforms in a PLL in locked condition, (b) calculation of phase error.

In the waveforms of Fig. 15.7(a), two quantities are unknown: ϕ_0 and the dc level of V_{cont}. To determine these values, we construct the VCO and PD characteristics [Fig. 15.7(b)]. If the input and output frequencies are equal to ω_1, then the required oscillator control voltage is unique and equal to V_1. This voltage must be produced by the phase detector, demanding a phase error determined by the PD characteristic. More specifically, since $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$ and $\overline{V_{PD}} = K_{PD}\Delta\phi$, we can write

$$V_1 = \frac{\omega_1 - \omega_0}{K_{VCO}}, \tag{15.3}$$

and

$$\phi_0 = \frac{V_1}{K_{PD}} \tag{15.4}$$

$$= \frac{\omega_1 - \omega_0}{K_{PD}K_{VCO}}. \tag{15.5}$$

Equation (15.5) reveals two important points: (1) as the input frequency of the PLL varies, so does the phase error; (2) to minimize the phase error, $K_{PD}K_{VCO}$ must be maximized.

Example 15.3

A PLL incorporates a VCO and a PD having the characteristics shown in Fig. 15.8. Explain what happens as the input frequency varies in the locked condition.

Solution

The PD characteristic is relatively linear near the origin but exhibits a small-signal gain of zero if the phase difference equals $\pm\pi/2$, at which point the average output is equal to $\pm V_0$. Now suppose

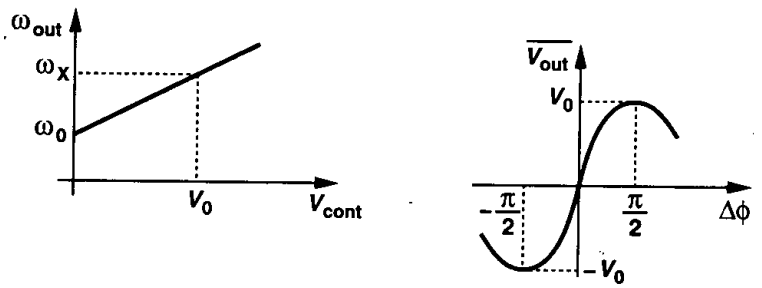


Figure 15.8

the input frequency increases from ω_0, requiring a greater control voltage. If the frequency is high enough ($= \omega_x$) to mandate $V_{cont} = V_0$, then the PD must operate at the peak of its characteristic. However, the PD gain drops to zero here and the feedback loop fails. Thus, the circuit cannot lock if $\omega_{in} = \omega_X$.

With the basic understanding of PLLs developed thus far, we now return to Eq. (15.2). The exact equality of the input and output frequencies of a PLL in the locked condition is a critical attribute. The significance of this property can be seen from two observations. First, in many applications, even a very small (deterministic) frequency error may prove unacceptable. For example, if a data stream is to be processed synchronously by a clocked system, even a slight difference between the data rate and the clock frequency results in a "drift," creating errors (Fig. 15.9). Second, the equality would *not* exist if the PLL compared the input and output frequencies rather than phases. As illustrated in Fig. 15.10(a), a loop employing a frequency detector (FD) would suffer from a finite difference between ω_{in} and ω_{out} due to various mismatches and other nonidealities. This can be understood by an analogy with the unity-gain feedback circuit of Fig. 15.10(b). Even if the op amp's

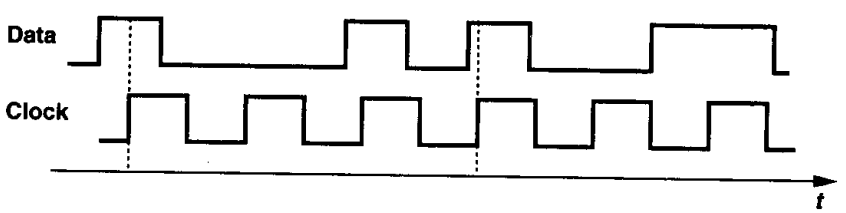
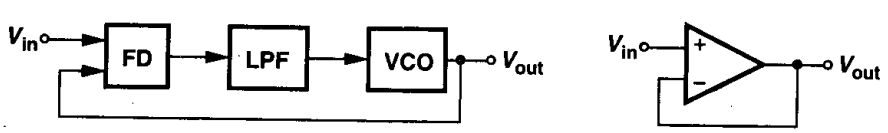


Figure 15.9 Drift of data with respect to clock in the presence of small frequency error.



(a) (b)

Figure 15.10 (a) Frequency-locked loop, (b) unity-gain feedback amplifier.

open-loop gain is infinity, the input-referred offset voltage leads to a finite error between V_{in} and V_{out}.

Small Transients in Locked Condition Let us now analyze the response of a PLL in locked condition to small phase or frequency transients at the input.

Consider a PLL in the locked condition and assume the input and output waveforms can be expressed as

$$V_{in}(t) = V_A \cos \omega_1 t \tag{15.6}$$

$$V_{out}(t) = V_B \cos(\omega_1 t + \phi_0), \tag{15.7}$$

where higher harmonics are neglected and ϕ_0 is the static phase error. Suppose, as shown in Fig. 15.11, the input experiences a phase step of ϕ_1 at $t = t_1$, i.e., $\phi_{in} = \omega_1 t + \phi_1 u(t - t_1)$.[2] Since the output of the LPF does not change instantaneously, the VCO initially continues to

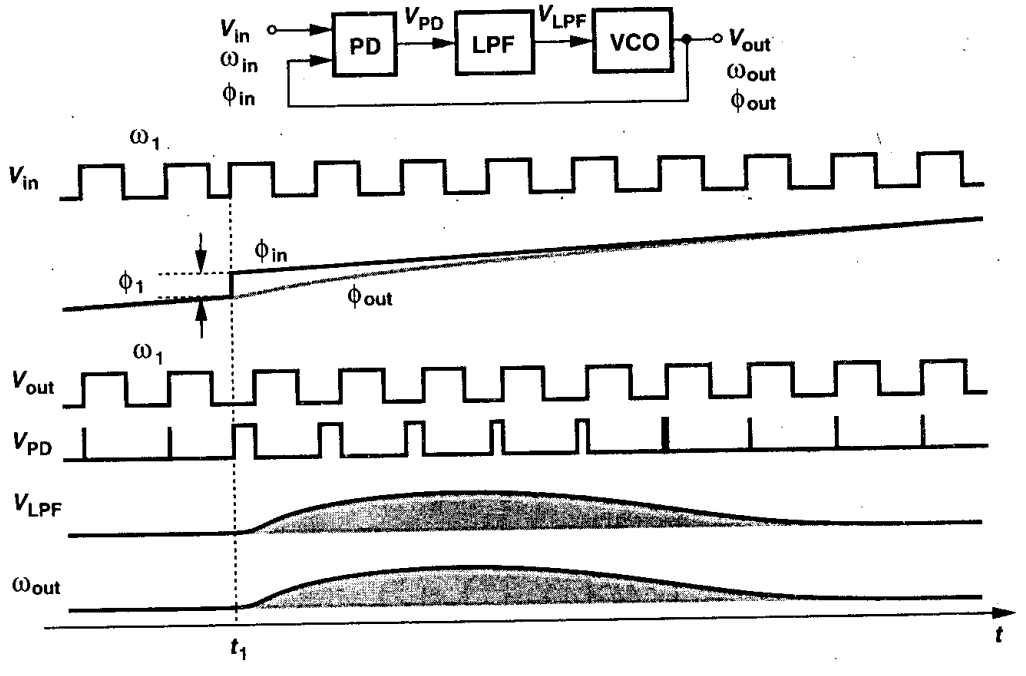


Figure 15.11 Response of a PLL to a phase step.

oscillate at ω_1. The growing phase difference between the input and the output then creates wide pulses at the output of the PD, forcing V_{LPF} to rise gradually. As a result, the VCO frequency begins to change, attempting to minimize the phase error. Note that the loop is not locked during the transient because the phase error varies with time.

[2] In this example, ϕ_{in} and ϕ_{out} denote the *total* phases of the input and output, respectively.

What happens after the VCO frequency begins to change? If the loop is to return to lock, ω_{out} must eventually go back to ω_1, requiring that V_{LPF} and hence $\phi_{out} - \phi_{in}$ also return to their original values. Since ϕ_{in} has changed by ϕ_1, the variation in the VCO frequency is such that the *area* under ω_{out} provides an additional phase of ϕ_1 in ϕ_{out}:

$$\int_{t1}^{\infty} \omega_{out} dt = \phi_1. \tag{15.8}$$

Thus, when the loop settles, the output becomes equal to

$$V_{out}(t) = V_B \cos[\omega_1 t + \phi_0 + \phi_1 u(t - t_1)]. \tag{15.9}$$

Consequently, as shown in Fig. 15.11, ϕ_{out} gradually "catches up" with ϕ_{in}.

It is important to make two observations. (1) After the loop returns to lock, *all* of the parameters (except for the total input and output phases) assume their original values. That is, $\phi_{in} - \phi_{out}$, V_{LPF}, and the VCO frequency remain unchanged—an expected result because these three parameters bear a one-to-one relationship and the input frequency has stayed the same. (2) The control voltage of the oscillator can serve as a suitable test point in the analysis of PLLs. While it is difficult to measure the time variations of phase and frequency in Fig. 15.11, $V_{cont}(= V_{LPF})$ can be readily monitored in simulations and measurements.

The reader may wonder whether an input phase step always gives rise to the response shown in Fig. 15.11. For example, is it possible for V_{LPF} to ring before settling to its final value? Such behavior is indeed possible and will be quantified in Section 15.1.3.

Let us now examine the response of PLLs to a small input frequency step $\Delta\omega$ at $t = t_1$ (Fig. 15.12). As with the case of a phase step, the VCO initially continues to oscillate at

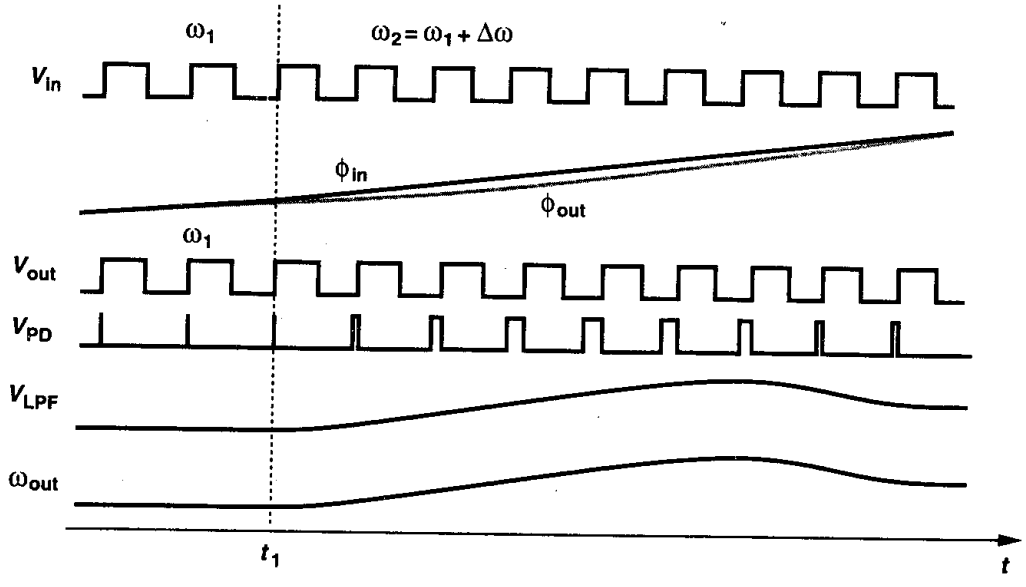


Figure 15.12 Response of a PLL to a small frequency step.

ω_1. Thus, the PD generates increasingly wider pulses, and V_{LPF} rises with time. As ω_{out} approaches $\omega_1 + \Delta\omega$, the width of the pulses generated by the PD decreases, eventually settling to a value that produces a dc component equal to $(\omega_1 + \Delta\omega - \omega_0)/K_{VCO}$. In contrast to the case of phase step, the response of a PLL to a frequency step entails a permanent change in both the control voltage and the phase error. If the input frequency is varied slowly, ω_{out} simply "tracks" ω_{in}.

The exact settling behavior of PLLs depends on the various loop parameters and will be studied in Section 15.1.3. But, to arrive at an important observation, we consider the phase step response depicted in Fig. 15.13, where V_{cont} rings before settling to its final value.

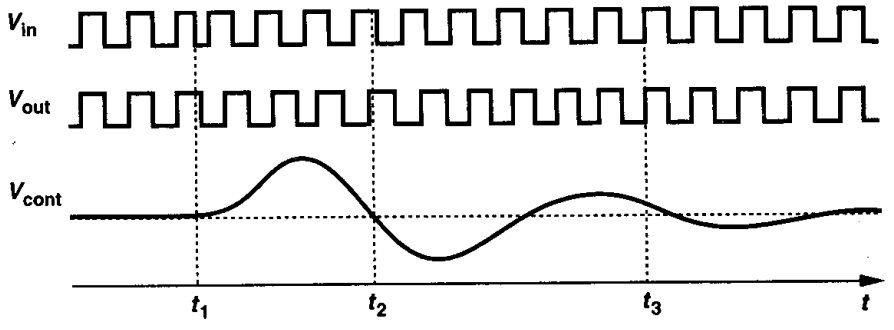


Figure 15.13 Example of phase step response.

Consider the state of the loop at $t = t_2$. At this point, the output frequency is equal to its final value (because V_{cont} is equal to its final value) but the loop continues the transient because the phase error deviates from the required value. Similarly, at $t = t_3$, the phase error is equal to its final value but the output frequency is not. In other words, for the loop to settle, both the phase and the frequency must settle to proper values.

Example 15.4

Consider the PLL shown in Fig. 15.14, where an external voltage V_{ex} is added to the output of the low-pass filter.[3] (a) Determine the phase error and V_{LPF} if the loop is locked and $V_{ex} = V_1$. (b) Suppose V_{ex} steps from V_1 to V_2 at $t = t_1$. How does the loop respond?

Solution

(a) If the loop is locked, $\omega_{out} = \omega_{in}$ and $V_{cont} = (\omega_{in} - \omega_0)/K_{VCO}$. Thus, $V_{LPF} = (\omega_{in} - \omega_0)/K_{VCO} - V_1$ and $\Delta\phi = V_{LPF}/K_{PD} = (\omega_{in} - \omega_0)/(K_{PD}K_{VCO}) - V_1/K_{PD}$.

(b) When V_{ex} steps from V_1 to V_2, V_{cont} immediately goes from $(\omega_{in} - \omega_0)/K_{VCO}$ to $(\omega_{in} - \omega_0)/K_{VCO} + (V_2 - V_1)$, changing the VCO frequency to $\omega_{in} - K_{VCO}(V_1 - V_2)$. Since V_{LPF} cannot change instantaneously, the PD begins to generate increasingly wider pulses, raising V_{LPF} and increasing ω_{out}. When the loop returns to lock, ω_{out} becomes equal to ω_{in} and $V_{LPF} = (\omega_{in} - \omega_0)/K_{VCO} - V_2$. The phase error also changes to $(\omega_{in} - \omega_0)/(K_{PD}K_{VCO}) - V_2/K_{PD}$. Note that the

[3]This topology is used for some types of frequency modulation in wireless communication.

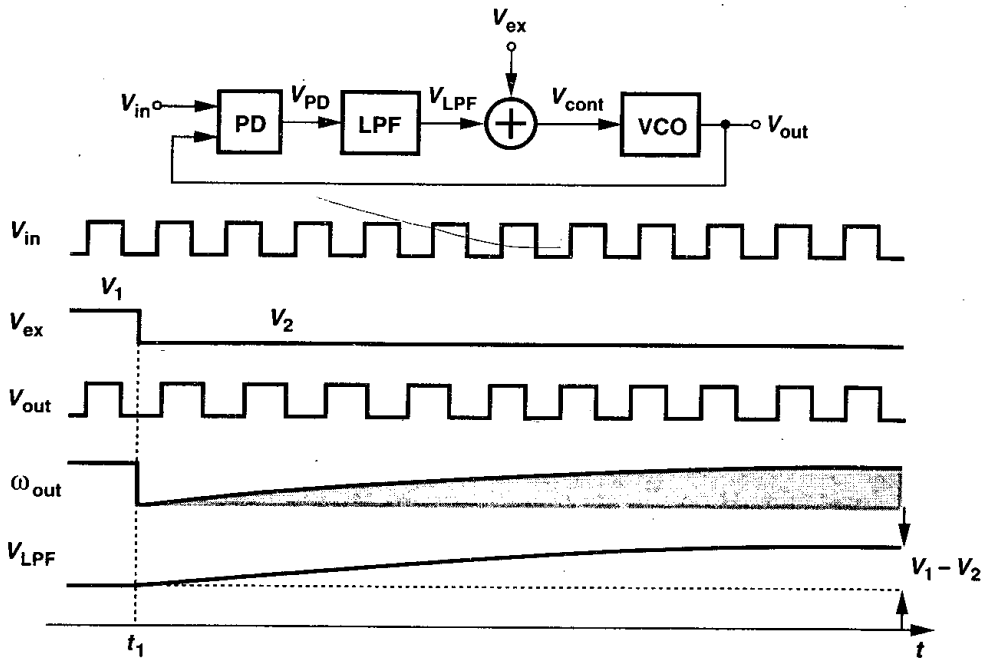


Figure 15.14

area under ω_{out} during the transient is equal to the change in the output phase and hence the change in the phase error:

$$\int_{t1}^{\infty} \omega_{out} dt = \frac{V_1 - V_2}{K_{PD}}. \tag{15.10}$$

From our study thus far, we conclude that phase-locked loops are "dynamic" systems, i.e., their response depends on the past values of the input and output. This is to be expected because the low-pass filter and the VCO introduce poles (and possibly zeros) in the loop transfer function. Moreover, we note that, so long as the input and the output remain perfectly periodic (i.e., $\phi_{in} = \omega_{in} t$ and $\phi_{out} = \omega_{in} t + \phi_0$), the loop operates in the steady state, exhibiting no transient. Thus, the PLL only responds to variations in the *excess* phase of the input or output. For example, in Fig. 15.11, $\phi_{in} = \omega_1 t + \phi_1 u(t - t_1)$ and in Fig. 15.12, $\phi_{in} = \omega_1 t + \Delta\omega \cdot t u(t - t_1)$.

15.1.3 Dynamics of Simple PLL

With the qualitative analysis of PLLs in the previous section, we can now study their transient behavior more rigorously. Assuming the loop is initially locked, we treat the PLL as a feedback system but recognize that the output quantity in this analysis must be

the (excess) phase of the VCO because the "error amplifier" can only compare phases. Our objective is to determine the transfer function $\Phi_{out}(s)/\Phi_{in}(s)$ for both open-loop and closed-loop systems and subsequently study the time-domain response. Note that the dimensions change from phase to voltage through the PD and from voltage to phase through the VCO.

What does $\Phi_{out}(s)/\Phi_{in}(s)$ signify? An analogy with more familiar transfer functions proves useful here. A circuit having a transfer function $V_{out}(s)/V_{in}(s) = 1/(1 + s/\omega_0)$ is considered a low-pass filter because if V_{in} varies rapidly, V_{out} cannot fully track the input variations. Similarly, $\Phi_{out}(s)/\Phi_{in}(s)$ reveals how the output phase tracks the input phase if the latter changes slowly or rapidly.

To visualize the variation of the excess phase with time, consider the waveforms in Fig. 15.15. The period varies slowly in Fig. 15.15(a) and rapidly in Fig. 15.15(b). Thus, $y_2(t)$ experiences faster phase variation than does $y_2(t)$.

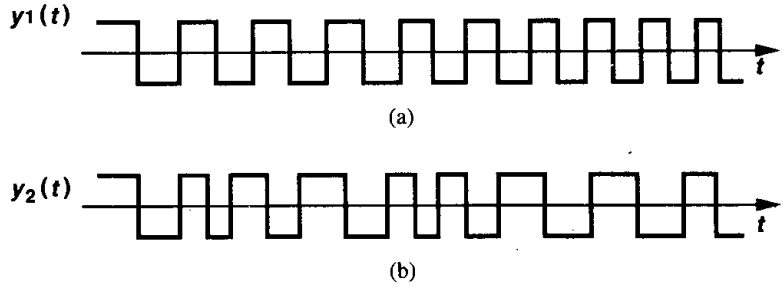


Figure 15.15 Slow and fast variation of the excess phase.

Let us construct a linear model of the PLL, assuming a first-order low-pass filter for simplicity. The PD output contains a dc component equal to $K_{PD}(\phi_{out} - \phi_{in})$ as well as high-frequency components. Since the latter are suppressed by the LPF, we simply model the PD by a subtractor whose output is "amplified" by K_{PD}. Illustrated in Fig. 15.16, the overall PLL model consists of the phase subtractor, the LPF transfer function $1/(1 + s/\omega_{LPF})$,

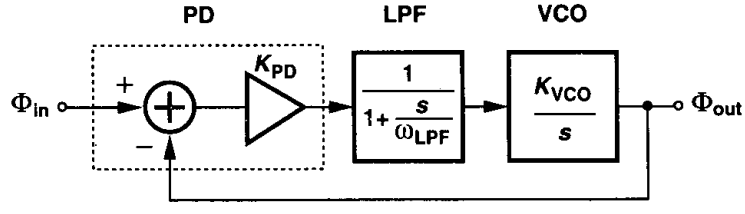


Figure 15.16 Linear model of type I PLL.

where ω_{LPF} denotes the -3-dB bandwidth, and the VCO transfer function K_{VCO}/s. Here, Φ_{in} and Φ_{out} denote the excess phases of the input and output waveforms, respectively. For example, if the total input phase experiences a step change, $\phi_1 u(t)$, then $\Phi_{in}(s) = \phi_1/s$.

The open-loop transfer function is given by

$$H(s)|_{\text{open}} = \frac{\Phi_{out}}{\Phi_{in}}(s)|_{\text{open}} \tag{15.11}$$

$$= K_{PD} \cdot \frac{1}{1 + \dfrac{s}{\omega_{LPF}}} \cdot \frac{K_{VCO}}{s}, \tag{15.12}$$

revealing one pole at $s = -\omega_{LPF}$ and another at $s = 0$. Note that the loop gain is equal to $H(s)|_{\text{open}}$ because of the unity feedback factor. Since the loop gain contains a pole at the origin, the system is called "type I."

Before computing the closed-loop transfer function, let us make an important observation. What is the loop gain if s is very small, i.e., if the input excess phase varies very slowly? Owing to the pole at the origin, the loop gain goes to infinity as s approaches zero, a point of contrast to the feedback circuits studied in Chapters 8 and 10. Thus, the phase-locked loop (under closed-loop, locked condition) ensures that the change in ϕ_{out} is *exactly* equal to the change in ϕ_{in} as s goes to zero. This result predicts two interesting properties of PLLs. First, if the input excess phase varies very slowly, the output excess phase "tracks" it. (After all, ϕ_{out} is "locked" to ϕ_{in}.) Second, if the transients in ϕ_{in} have decayed (another case corresponding to $s \to 0$), then the change in ϕ_{out} is precisely equal to the change in ϕ_{in}. This is indeed true in the example depicted in Fig. 15.11.

From (15.12), we can write the closed-loop transfer function as:

$$H(s)|_{\text{closed}} = \frac{K_{PD}K_{VCO}}{\dfrac{s^2}{\omega_{LPF}} + s + K_{PD}K_{VCO}}. \tag{15.13}$$

For the sake of brevity, we hereafter denote $H(s)|_{\text{closed}}$ simply by $H(s)$ or Φ_{out}/Φ_{in}. As expected, if $s \to 0$, $H(s) \to 1$ because of the infinite loop gain.

In order to analyze $H(s)$ further, we derive a relationship that allows a more intuitive understanding of the system. Recall that the instantaneous frequency of a waveform is equal to the time derivative of the phase: $\omega = d\phi/dt$. Since the frequency and the phase are related by a linear operator, the transfer function of (15.13) applies to variations in the input and output frequencies as well:

$$\frac{\omega_{out}}{\omega_{in}}(s) = \frac{K_{PD}K_{VCO}}{\dfrac{s^2}{\omega_{LPF}} + s + K_{PD}K_{VCO}}. \tag{15.14}$$

For example, this result predicts that if ω_{in} changes very slowly ($s \to 0$), then ω_{out} tracks ω_{in}, again an expected result because the loop is assumed locked. Equation (15.14) also indicates that if ω_{in} changes abruptly but the system is given enough time to settle ($s \to 0$), then the change in ω_{out} equals that in ω_{in} (as illustrated in the example of Fig. 15.12).

The above observation aids the analysis in two directions. First, some transient responses of the closed-loop system may be simpler to visualize in terms of changes in the frequency quantities rather than phase quantities. Second, since a change in ω_{out} must be accompanied

by a change in V_{cont}, we have

$$H(s) = K_{VCO} \cdot \frac{V_{cont}}{\omega_{in}}(s). \tag{15.15}$$

That is, monitoring the response of V_{cont} to variations in ω_{in} indeed yields the response of the closed-loop system.

The second-order transfer function of (15.13) suggests that the step response of the type I system can be overdamped, critically damped, or underdamped. To derive the condition for each case, we rewrite the denominator in a familiar form used in control theory, $s^2+2\zeta\omega_n s+\omega_n^2$, where ζ is the "damping ratio" and ω_n is the "natural frequency." That is,

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{15.16}$$

where

$$\omega_n = \sqrt{\omega_{LPF} K_{PD} K_{VCO}} \tag{15.17}$$

$$\zeta = \frac{1}{2}\sqrt{\frac{\omega_{LPF}}{K_{PD}K_{VCO}}}. \tag{15.18}$$

The two poles of the closed-loop system are given by

$$s_{1,2} = -\zeta\omega_n \pm \sqrt{(\zeta^2 - 1)\omega_n^2} \tag{15.19}$$

$$= (-\zeta \pm \sqrt{\zeta^2 - 1})\omega_n. \tag{15.20}$$

Thus, if $\zeta > 1$, both poles are real, the system is overdamped, and the transient response contains two exponentials with time constants $1/s_1$ and $1/s_2$. On the other hand, if $\zeta < 1$, the poles are complex and the response to an input frequency step $\omega_{in} = \Delta\omega u(t)$ is equal to

$$\omega_{out}(t) = \left\{1 - e^{-\zeta\omega_n t}[\cos(\omega_n\sqrt{1-\zeta^2}t) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin(\omega_n\sqrt{1-\zeta^2}t)]\right\}\Delta\omega u(t) \tag{15.21}$$

$$= [1 - \frac{1}{\sqrt{1-\zeta^2}}e^{-\zeta\omega_n t}\sin(\omega_n\sqrt{1-\zeta^2}t + \theta)]\Delta\omega u(t), \tag{15.22}$$

where ω_{out} denotes the change in the output frequency and $\theta = \sin^{-1}\sqrt{1-\zeta^2}$. Thus, as shown in Fig. 15.17, the step response contains a sinusoidal component with a frequency $\omega_n\sqrt{1-\zeta^2}$ that decays with a time constant $(\zeta\omega_n)^{-1}$. Note that the system exhibits the same response if a phase step is applied to the input and the output phase is observed.

The settling speed of PLLs is of great concern in most applications. Equation (15.22) indicates that the exponential decay determines how fast the output approaches its final

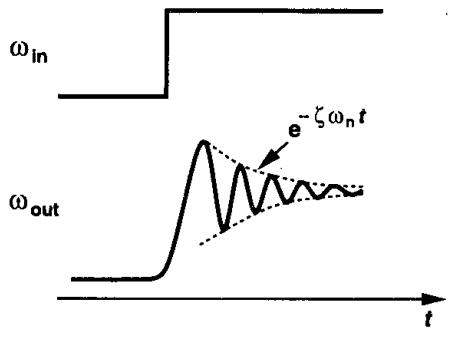


Figure 15.17 Underdamped response of PLL to a frequency step.

value, implying that $\zeta\omega_n$ must be maximized. For the type I PLL under study here, (15.17) and (15.18) yield

$$\zeta\omega_n = \frac{1}{2}\omega_{LPF}. \tag{15.23}$$

This result reveals a critical trade-off between the settling speed and the ripple on the VCO control line: the lower ω_{LPF}, the greater the suppression of the high-frequency components produced by the PD but the longer the settling time constant.

Example 15.5

A cellular telephone incorporates a 900-MHz phase-locked loop to generate the carrier frequencies. If $\omega_{LPF} = 2\pi \times (20 \text{ kHz})$ and the output frequency is to be changed from 901 MHz to 901.2 MHz, how long does the PLL output frequency take to settle within 100 Hz of its final value?

Solution

Since the step size is 200 kHz, we have

$$[1 - e^{-\zeta\omega_n t_s}\sin(\omega_n\sqrt{1-\zeta^2}t_s + \theta)] \times 200 \text{ kHz} = 200 \text{ kHz} - 100 \text{ Hz}. \tag{15.24}$$

Thus,

$$e^{-\zeta\omega_n t_s}\sin(\omega_n\sqrt{1-\zeta^2}t_s + \theta) = \frac{100 \text{ Hz}}{200 \text{ kHz}}. \tag{15.25}$$

In the worst case, the sinusoid is equal to unity and

$$e^{-\zeta\omega_n t_s} = 0.0005. \tag{15.26}$$

That is,

$$t_s = \frac{7.6}{\zeta\omega_n} \tag{15.27}$$

$$= \frac{15.2}{\omega_{LPF}} \tag{15.28}$$

$$= 0.12 \text{ ms}. \tag{15.29}$$

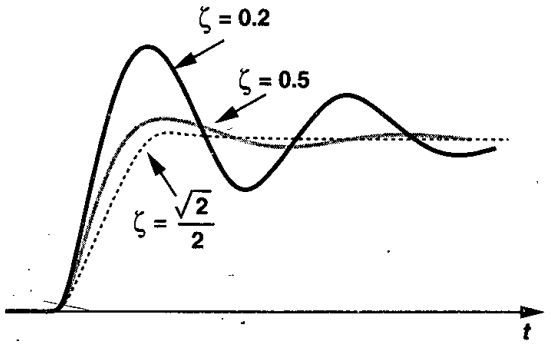


Figure 15.18 Underdamped response of a second-order system for various values of ζ.

In addition to the product $\zeta\omega_n$, the value of ζ itself is also important. Illustrated in Fig. 15.18 for several values of ζ and a constant ω_n, the step response exhibits severe ringing for $\zeta < 0.5$. In view of process and temperature variation of the loop parameters, ζ is usually chosen to be greater than $\sqrt{2}/2$ or even 1 to avoid excessive ringing.[4]

The choice of ζ entails other trade-offs as well. First, (15.18) implies that as ω_{LPF} is reduced to minimize the ripple on the control voltage, the stability degrades. Second, (15.5) and (15.18) indicate that both the phase error and ζ are inversely proportional to $K_{PD}K_{VCO}$; lowering the phase error inevitably makes the system less stable. In summary, the type I PLL suffers from trade-offs between the settling speed, the ripple on the control voltage (i.e., the quality of the output signal), the phase error, and the stability.

The stability behavior of PLLs can also be analyzed graphically, providing more insight. Recall from Chapter 10 that the Bode plots of the magnitude and phase of the loop gain readily yield the phase margin. Let us utilize (15.12) to construct such plots. As shown in Fig. 15.19, the loop gain begins from infinity at $\omega = 0$ and falls at a rate of 20 dB/dec for $\omega < \omega_{LPF}$ and at a rate of 40 dB/dec thereafter. The phase begins at $-90°$ and asymptotically reaches $-180°$.

What happens if a higher $K_{PD}K_{VCO}$ is chosen so as to minimize $\phi_{out} - \phi_{in}$? Since the entire gain plot in Fig. 15.19 is shifted up, the gain crossover moves to the right, thus degrading the phase margin. This is consistent with the dependence of ζ upon $K_{PD}K_{VCO}$.

As observed thus far, $K_{PD}K_{VCO}$ impacts many important parameters of PLLs. This quantity is sometimes called the loop gain (even though it is not dimensionless) due to the resemblance of $\Delta\phi = (\omega_{out} - \omega_0)/(K_{PD}K_{VCO})$ to the error equation in a feedback system.

The stability behavior of type I PLLs can also be analyzed by the locus of their poles in the complex plane as the parameter $K_{PD}K_{VCO}$ varies (Fig. 15.20). With $K_{PD}K_{VCO} = 0$,

[4]The value of ζ may also yield peaking in the transfer function. Thus, some applications require a ζ of 5 to 10 to avoid peaking in the presence of higher order poles.

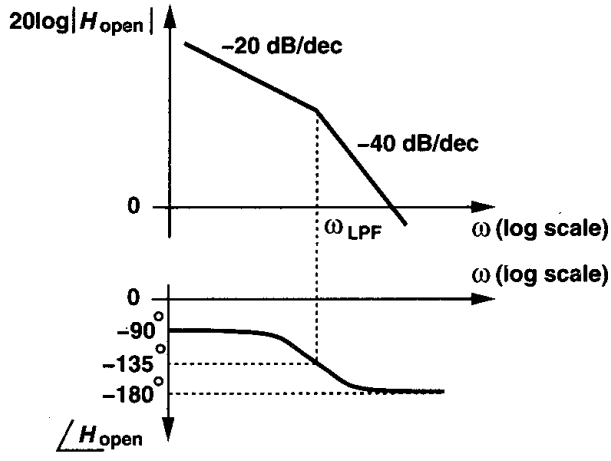


Figure 15.19 Bode plots of type I PLL.

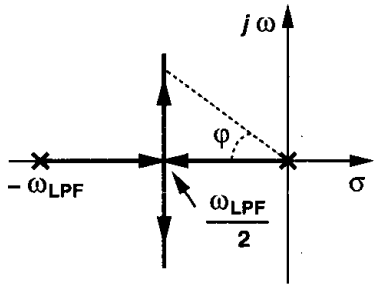


Figure 15.20 Root locus of type I PLL.

the loop is open, $\zeta = \infty$, and the two poles are given by $s_1 = -\omega_{LPF}$ and $s_2 = 0$. As $K_{PD}K_{VCO}$ increases (i.e., the feedback becomes stronger), ζ drops and the two poles, given by $s_{1,2} = (-\zeta \pm \sqrt{\zeta^2 - 1})\omega_n$, move toward each other on the real axis. For $\zeta = 1$ (i.e., $K_{PD}K_{VCO} = \omega_{LPF}/4$), $s_1 = s_2 = -\zeta\omega_n = -\omega_{LPF}/2$. As $K_{PD}K_{VCO}$ increases further, the two poles become complex, with a real part equal to $-\zeta\omega_n = -\omega_{LPF}/2$, moving in parallel with the $j\omega$ axis.

We recognize from Fig. 15.20 that, as s_1 and s_2 move away from the real axis, the system becomes less stable. In fact, the reader can prove that $\cos\varphi = \zeta$ (Problem 15.8), concluding that as φ approaches 90°, ζ drops to zero.

Another transfer function that reveals the settling behavior of PLLs is that of the error at the output of the phase subtractor in Fig. 15.16. Defined as $H_e(s) = (\phi_{in} - \phi_{out})/\phi_{in}$, this transfer function can be obtained by noting that $\phi_{out}/\phi_{in} = H(s)$ and, from (15.13),

$$H_e(s) = 1 - H(s) \tag{15.30}$$

$$= \frac{s^2 + 2\zeta\omega_n s}{s^2 + 2\zeta\omega_n s + \omega_n}. \tag{15.31}$$

As expected, $H_e(s) \to 0$ if $s \to 0$ because the output tracks the input when the input varies very slowly or the transient has settled.

Example 15.6

Suppose a type I PLL experiences a frequency step $\Delta\omega$ at $t = 0$. Calculate the change in the phase error.

Solution

The Laplace transform of the frequency step equals $\Delta\omega/s$. Since $H_e(s)$ relates the phase error to the input phase, we write $\Phi_{in}(s) = (\Delta\omega/s)/s = \Delta\omega/s^2$. Thus, the Laplace transform of the phase error is

$$\Phi_e(s) = H_e(s) \cdot \frac{\Delta\omega}{s^2} \tag{15.32}$$

$$= \frac{s^2 + 2\zeta\omega_n s}{s^2 + 2\zeta\omega_n s + \omega_n^2} \cdot \frac{\Delta\omega}{s^2}. \tag{15.33}$$

From the final value theorem,

$$\phi_e(t = \infty) = \lim_{s \to 0} s\Phi_e(s) \tag{15.34}$$

$$= \frac{2\zeta}{\omega_n}\Delta\omega \tag{15.35}$$

$$= \frac{\Delta\omega}{K_{PD}K_{VCO}}, \tag{15.36}$$

which agrees with (15.5).

15.2 Charge-Pump PLLs

While type I PLLs have been realized widely in discrete form, their shortcomings often prohibit usage in high-performance integrated circuits. In addition to the trade-offs between ζ, ω_{LPF}, and the phase error, type I PLLs suffer from another critical drawback: limited acquisition range.

15.2.1 Problem of Lock Acquisition

Suppose when a PLL circuit is turned on, its oscillator operates at a frequency far from the input frequency, i.e., the loop is not locked. Under what conditions does the loop "acquire" lock? The transition of the loop from unlocked to locked condition is a very nonlinear phenomenon because the phase detector senses unequal frequencies. The problem of lock acquisition in type I PLLs has been studied extensively [1, 2], but we state without proof

that the "acquisition range"[5] is on the order of ω_{LPF}, that is, the loop locks only if the difference between ω_{in} and ω_{out} is less than roughly ω_{LPF}.[6]

The problem of lock acquisition further tightens the trade-offs in type I PLLs. If ω_{LPF} is reduced to suppress the ripple on the control voltage, the acquisition range decreases. Note that even if the input frequency has a precisely controlled value, a wide acquisition range is often necessary because the VCO center frequency may vary considerably with process and temperature. In most of today's applications, the acquisition range of the simple PLL studied thus far proves inadequate.

In order to remedy the acquisition problem, modern PLLs incorporate frequency detection in addition to phase detection. Called "aided acquisition" and illustrated in Fig. 15.21, the idea is to compare ω_{in} and ω_{out} by means of a frequency detector, generate a dc component V_{LPF2} proportional to $\omega_{in} - \omega_{out}$, and apply the result to the VCO in a negative-feedback loop. At the beginning, the FD drives ω_{out} toward ω_{in} while the PD output remains "quiet." When $|\omega_{out} - \omega_{in}|$ is sufficiently small, the phase-locked loop takes over, acquiring lock. Such a scheme increases the acquisition range to the tuning range of the VCO.[7]

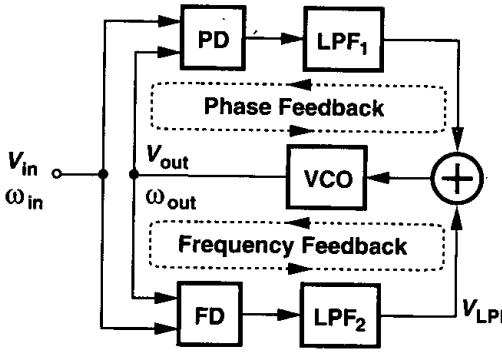


Figure 15.21 Addition of frequency detection to increase the acquisition range.

15.2.2 Phase/Frequency Detector and Charge Pump

For periodic signals, it is possible to merge the two loops of Fig. 15.21 by devising a circuit that can detect both phase and frequency differences. Called a phase/frequency detector (PFD) and illustrated conceptually in Fig. 15.22, the circuit employs sequential logic to create three states and respond to the rising (or falling) edges of the two inputs. If initially $Q_A = Q_B = 0$, then a rising transition on A leads to $Q_A = 1$, $Q_B = 0$. The circuit remains

[5] Acquisition range, tracking range, lock range, capture range, and pull-in range are often used to describe the behavior of PLLs in the presence of input or VCO frequency variation. For our purposes, the acquisition range, the capture range, and the pull-in range are the same. The tracking range refers to the input frequency range across which a locked PLL can track the input. With the addition of frequency detection, the acquisition range becomes equal to the tracking range (for periodic signals).

[6] This is a very rough estimate. In practice, the acquisition range may be several times narrower or wider. It is also assumed that the tuning range of the VCO is large enough not to limit the acquisition range.

[7] This may not be true if the input is not periodic.

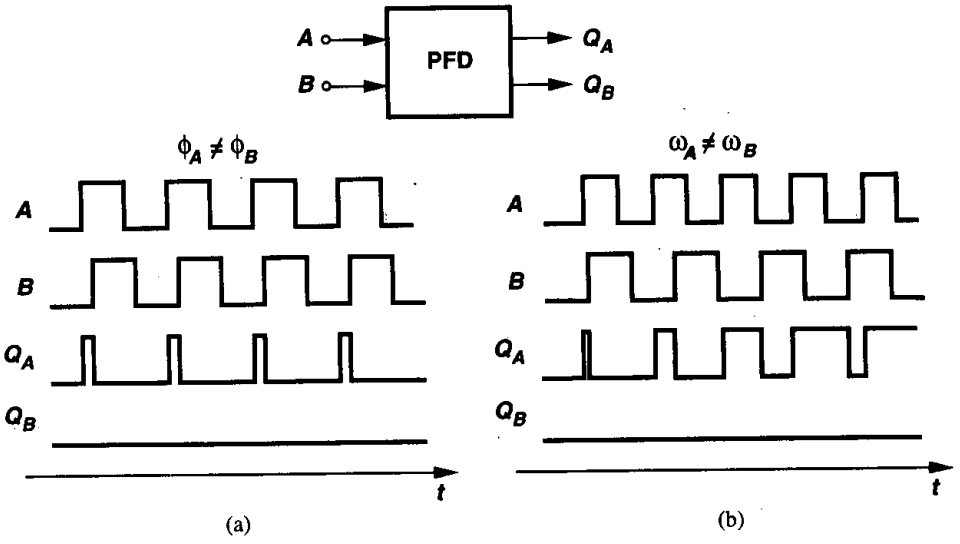


Figure 15.22 Conceptual operation of a PFD.

in this state until B goes high, at which point Q_A returns to zero. The behavior is similar for the B input.

In Fig. 15.22(a), the two inputs have equal frequencies but A leads B. The output Q_A continues to produce pulses whose width is proportional to $\phi_A - \phi_B$ while Q_B remains at zero. In Fig. 15.22(b), A has a higher frequency than B and Q_A generates pulses while Q_B does not. By symmetry, if A lags B or has a lower frequency than B, then Q_B produces pulses and Q_A remains quiet. Thus, the dc contents of Q_A and Q_B provide information about $\phi_A - \phi_B$ or $\omega_A - \omega_B$. The outputs Q_A and Q_B are called the "UP" and "DOWN" pulses, respectively.

Example 15.7

Explain whether a master-slave D flipflop can operate as a phase detector or a frequency detector. Assume the flipflop provides differential outputs.

Solution

As shown in Fig. 15.23(a), we first apply inputs having equal frequencies and a finite phase difference, assuming the output changes on the rising edge of the clock input. If A leads B, then V_{out} remains at a logical ONE indefinitely because the flipflop continues to sample the high levels of A. Conversely, if A lags B, then V_{out} remains low. Plotted in Fig. 15.23(b), the input-output characteristic of the circuit displays a very high gain at $\Delta\phi = 0, \pm\pi, \cdots$ and a zero gain at other values of $\Delta\phi$. The D flipflop is sometimes called a "bang-bang" phase detector to emphasize that the average value of V_{out} jumps from $-V_1$ to $+V_1$ as $\Delta\phi$ varies from slightly below zero to slightly above zero.

Now let us assume unequal frequencies for A and B. If the flipflop is to behave as a frequency detector, then the average value of V_{out} must exhibit different polarities for $\omega_A > \omega_B$ and $\omega_A < \omega_B$. However, as illustrated in Fig. 15.23(c), the average value is zero in both cases.

The circuit of Fig. 15.22 can be realized in various forms. Figure 15.24(a) shows a simple implementation consisting of two edge-triggered, resettable D flipflops with their D inputs tied to a logical ONE. The inputs of interest, A and B, serve as the clocks of the flipflops. If $Q_A = Q_B = 0$ and A goes high, Q_A rises. If this event is followed by a rising transition on B, Q_B goes high and the AND gate resets both flipflops. In other words, Q_A and Q_B are simultaneously high for a short time but the difference between their average values still represents the input phase or frequency difference correctly. Each flipflop can be implemented as shown in Fig. 15.24(b), where two RS latches are cross-coupled. Latch 1 and Latch 2 respond to the rising edges of CK and Reset, respectively.

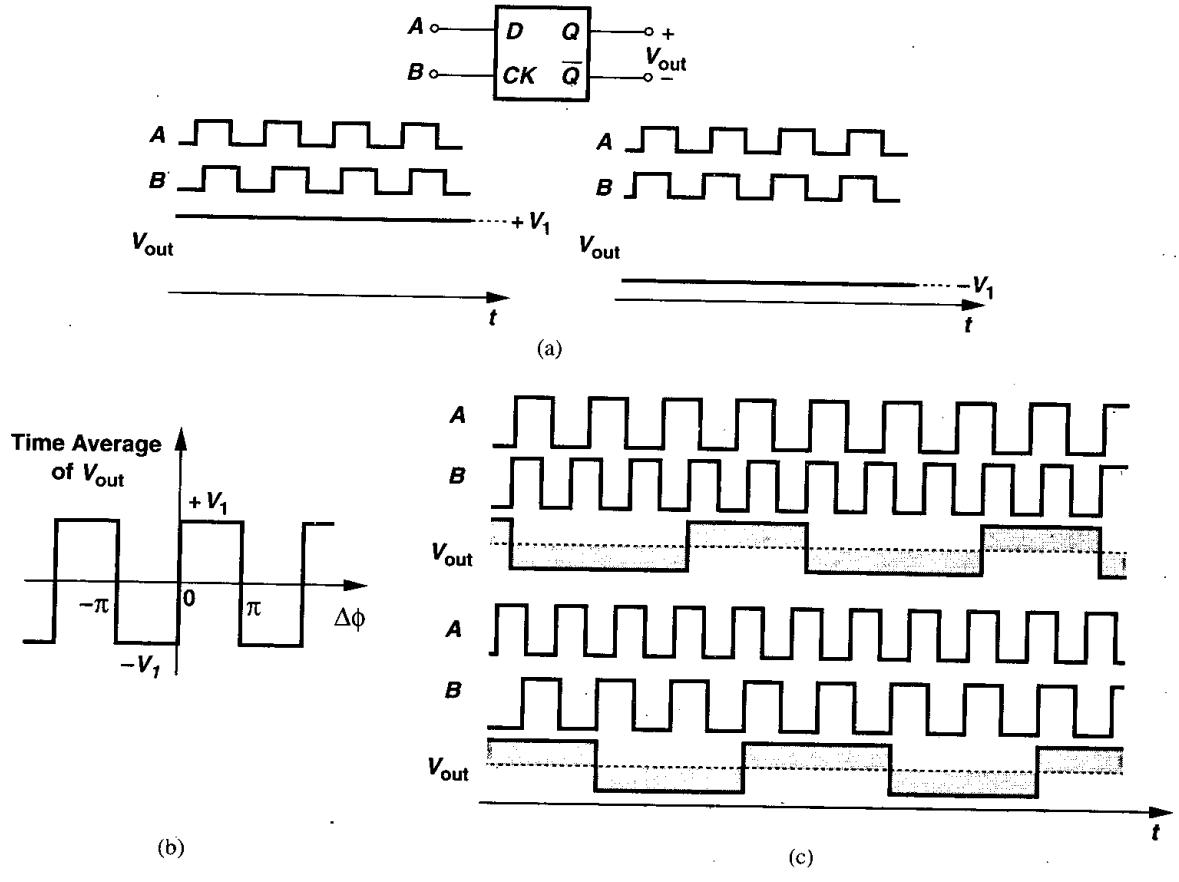
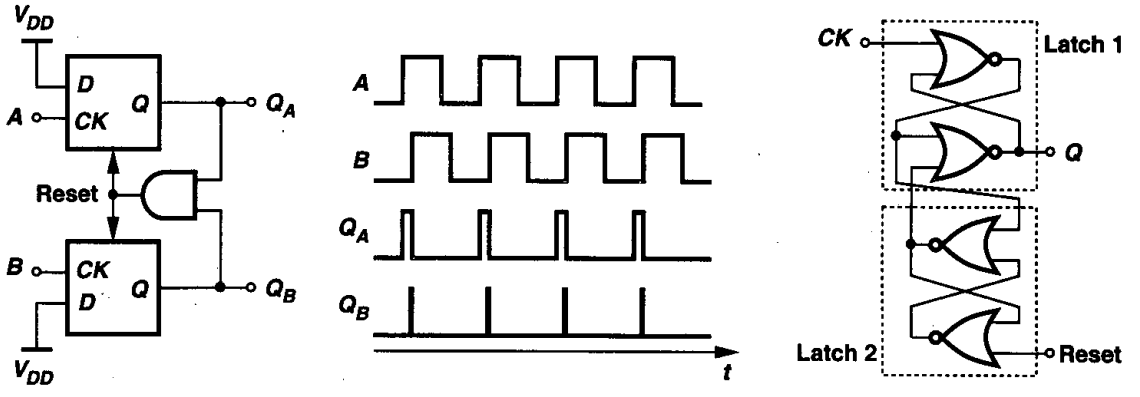


Figure 15.23 (a) D flipflop as a phase detector, (b) input/output characteristic, (c) response of D flipflop to unequal input frequencies.



(a) (b)

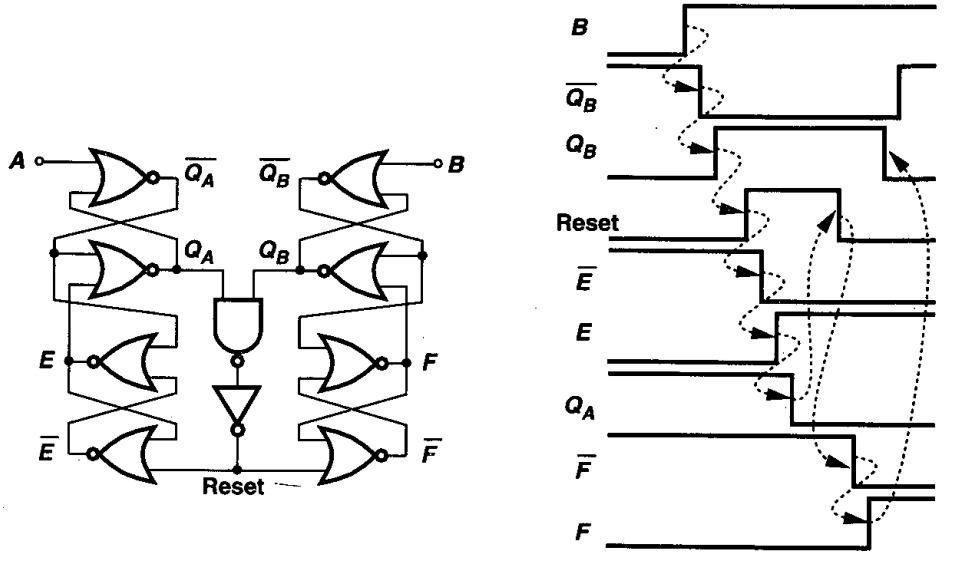
Figure 15.24 (a) Implementation of PFD, (b) implementation of D flipflop.

Example 15.8

Determine the width of the narrow reset pulses that appear in the Q_B waveform in Fig. 15.24(a).

Solution

Figure 15.25(a) illustrates the overall PFD at the gate level. If the circuit begins with $A = 1$, $Q_A = 1$, and $Q_B = 0$, a rising edge on B forces $\overline{Q}_B$ to go low and, one gate delay later, Q_B to go high. As



(a) (b)

Figure 15.25

shown in Fig. 15.25(b), this transition propagates to Reset, $\overline{E}$, E, Q_A, Reset, $\overline{F}$, F, and Q_B. Thus, the width of the pulse on Q_B is approximately equal to 10 gate delays.[8]

It is instructive to plot the input-output characteristic of the above PFD. Defining the output as the difference between the average values of Q_A and Q_B when $\omega_A = \omega_B$ and neglecting the effect of the narrow reset pulses, we note that the output varies symmetrically as $|\Delta\phi|$ begins from zero (Fig. 15.26). For $\Delta\phi = \pm 360°$, V_{out} reaches its maximum or minimum and subsequently changes sign.

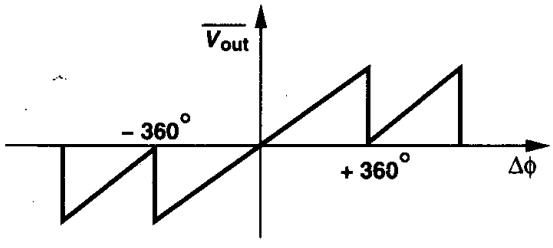


Figure 15.26 Input-output characteristic of the three-state PFD.

How is the PFD of Fig. 15.24(a) utilized in a phase-locked loop? Since the difference between the average values of Q_A and Q_B is of interest, the two outputs can be low-pass filtered and sensed differentially (Fig. 15.27). However, a more common approach is to interpose a "charge pump" (CP) between the PFD and the loop filter.

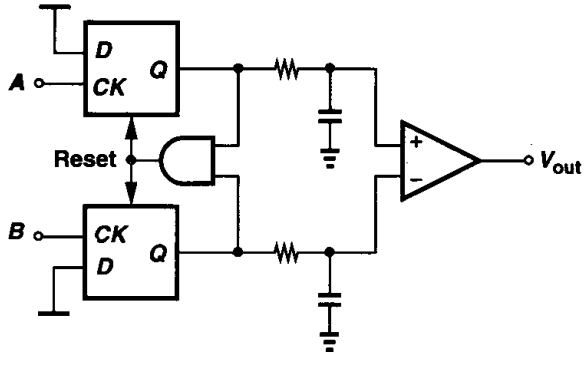


Figure 15.27 PFD followed by low-pass filters.

A charge pump consists of two switched current sources that pump charge into or out of the loop filter according to two logical inputs. Figure 15.28 illustrates a charge pump driven by a PFD and driving a capacitor. The circuit has three states. If $Q_A = Q_B = 0$, then S_1

[8]This is a rough approximation because the NAND gate, the inverter, and the NOR gates have different delays and fanouts.

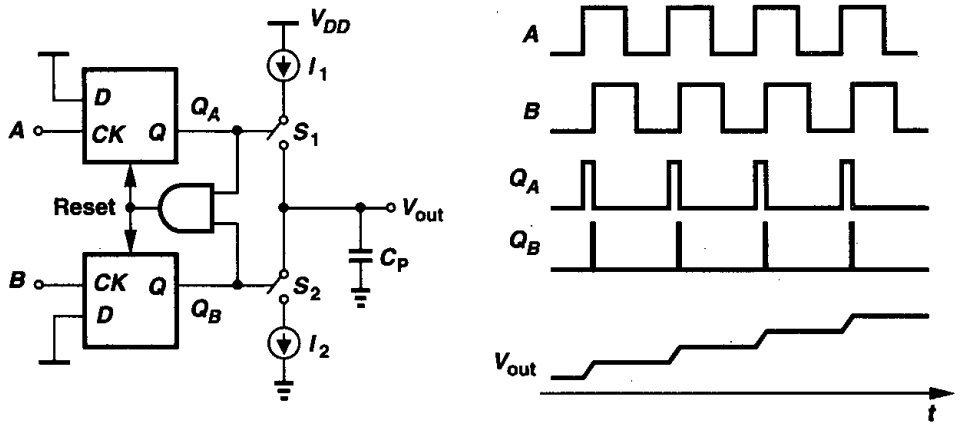


Figure 15.28 PFD with charge pump.

and S_2 are off and V_{out} remains constant. If Q_A is high and Q_B is low, then I_1 charges C_P. Conversely, if Q_A is low and Q_B is high, then I_2 discharges C_P. Thus, if, for example, A leads B, then Q_A continues to produce pulses and V_{out} rises steadily. Called UP and DOWN currents, respectively, I_1 and I_2 are nominally equal.

Example 15.9

What is the effect of the narrow pulses that appear in the Q_B waveform in Fig. 15.28?

Solution

Since Q_A and Q_B are simultaneously high for a finite period (approximately 10 gate delays from Example 15.8), the current supplied by the charge pump to C_P is affected. In fact, if $I_1 = I_2$, the current through S_1 simply flows through S_2 during the narrow reset pulse, leaving no current to charge C_P. Thus, as shown in Fig. 15.29, V_{out} remains constant after Q_B goes high.

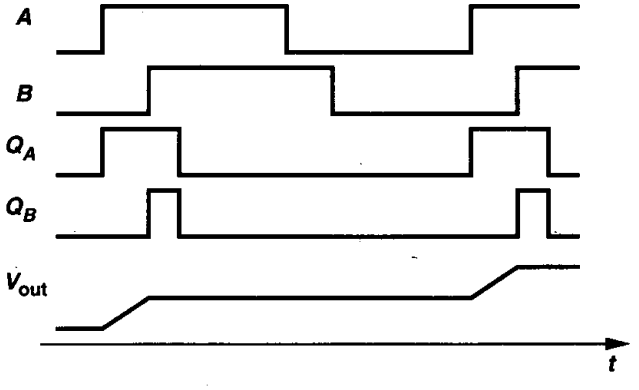


Figure 15.29

The circuit of Fig. 15.28 has an interesting property. If A, say, leads B by a finite amount, Q_A produces pulses indefinitely, allowing the charge pump to inject I_1 into C_P and forcing V_{out} to rise steadily. In other words, for a finite input error, the output eventually goes to $+\infty$ or $-\infty$, i.e., the "gain" of the circuit is infinity. The consequences of infinite gain are described below.

15.2.3 Basic Charge-Pump PLL

Let us now construct a PLL using the circuit of Fig. 15.28. Shown in Fig. 15.30 and called a charge-pump PLL, such an implementation senses the transitions at the input and output,

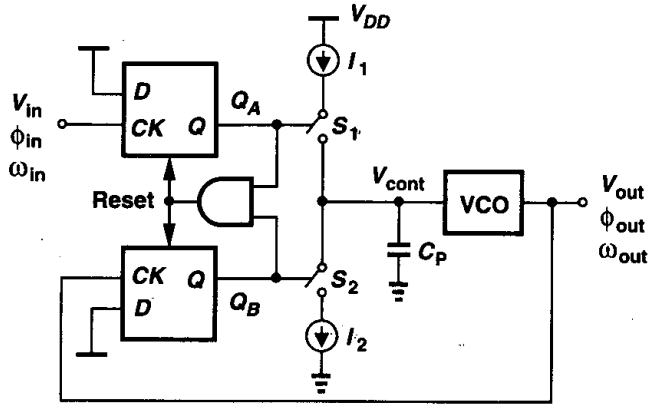


Figure 15.30 Simple charge-pump PLL.

detects phase or frequency differences, and activates the charge pump accordingly. When the loop is turned on, ω_{out} may be far from ω_{in}, and the PFD and the charge pump vary the control voltage such that ω_{out} approaches ω_{in}. When the input and output frequencies are sufficiently close, the PFD operates as a phase detector, performing phase lock. The loop locks when the phase difference drops to zero and the charge pump remains relatively idle.

As observed above, the gain of the PFD/CP combination is infinite, i.e., a nonzero (deterministic) difference between ϕ_{in} and ϕ_{out} leads to indefinite charge buildup on C_P. What is the consequence of this attribute in a charge-pump PLL? When the loop of Fig. 15.30 is locked, V_{cont} is finite. Therefore, the input phase error must be exactly *zero*.[9] This is in contrast to the behavior of the type I PLL, in which the phase error is finite and a function of the output frequency.

To gain more insight into the operation of the PLL shown in Fig. 15.30, let us ignore the narrow reset pulses on Q_A and Q_B and assume that after $\phi_{out} - \phi_{in}$ drops to zero, the PFD simply produces $Q_A = Q_B = 0$. The charge pump thus remains idle and C_P sustains a constant control voltage. Does this mean that the PFD and the CP are no longer needed?! If V_{cont} remains constant for a long time, the VCO frequency and phase begin to

[9] As explained in Section 15.3.1, mismatches still yield a finite phase error.

drift. In particular, the noise sources in the VCO create random variations in the oscillation frequency that can result in a large accumulation of phase error. The PFD then detects the phase difference, producing a corrective pulse on Q_A or Q_B that adjusts the VCO frequency through the charge pump and the filter. This is why we stated earlier that the PLL responds only to the *excess* phase of waveforms. We also note that, since in Fig. 15.30 phase comparison is performed in every cycle, the VCO phase and frequency cannot drift substantially.

Dynamics of CPPLL In order to quantify the behavior of charge-pump PLLs, we must develop a linear model for the combination of the PFD, the charge pump, and the low-pass filter, thereby obtaining the transfer function. We therefore raise two questions: (1) Is the PFD/CP/LPF combination in Fig. 15.28 a linear system? (2) If so, how can its transfer function be computed?

To answer the first question, we test the system for linearity. For example, as illustrated in Fig. 15.31(a), we double the input phase difference and see if V_{out} exactly doubles.

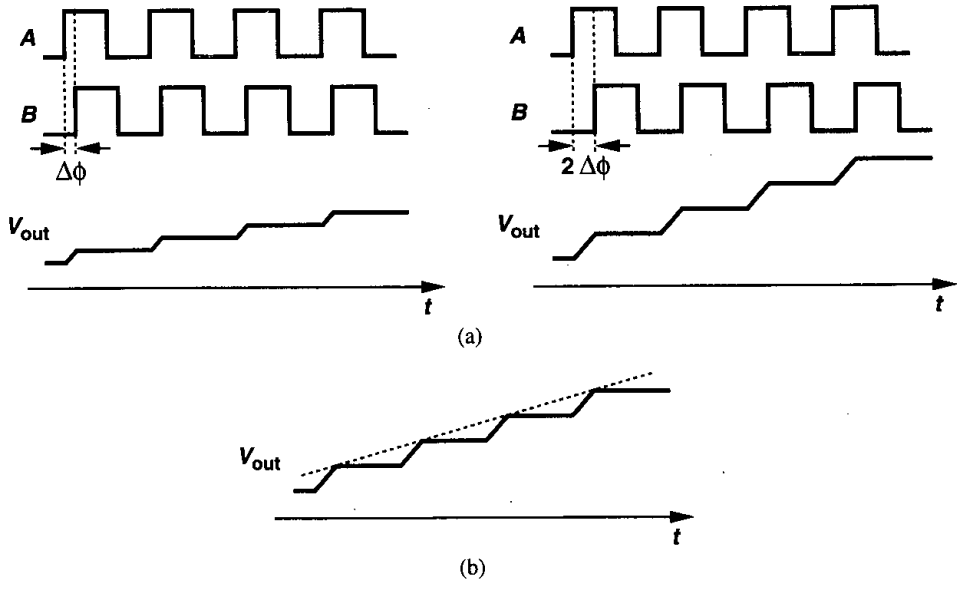


Figure 15.31 (a) Test of linearity of PFD/CP/LPF combination, (b) ramp approximation of the response.

Interestingly, the flat sections of V_{out} double but not the ramp sections. After all, the current charging or discharging C_P is constant, yielding a constant slope for the ramp—an effect similar to slewing in op amps. Thus, the system is not linear in the strict sense. To overcome this quandary, we approximate the output waveform by a ramp [Fig. 15.31(b)], arriving at a linear relationship between V_{out} and $\Delta\phi$. In a sense, we approximate a discrete-time system by a continuous-time model.

To answer the second question, we recall that the transfer function is the Laplace transform of the impulse response, requiring that we apply a phase difference impulse and

compute V_{out} in the time domain. Since a phase difference impulse is difficult to visualize, we apply a phase difference step, obtain V_{out}, and differentiate the result with respect to time.

Let us assume the input period is T_{in} and the charge pump provides a current of $\pm I_P$ to the capacitor. As shown in Fig. 15.32, we begin with a zero phase difference and, at $t = 0$, step the phase of B by ϕ_0, i.e., $\Delta\phi = \phi_0 u(t)$. As a result, Q_A or Q_B continues to produce

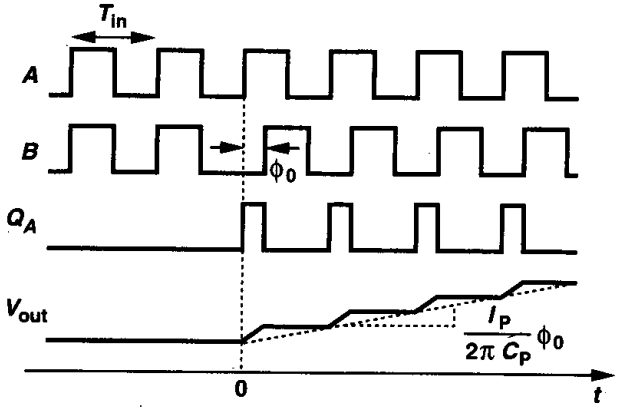


Figure 15.32 Step response of PFD/CP/LPF combination.

pulses that are $\phi_0 T_{in}/(2\pi)$ seconds wide, raising the output voltage by $(I_P/C_P)\phi_0 T_{in}/(2\pi)$ in every period.[10] Approximated by a ramp, V_{out} thus exhibits a slope of $(I_P/C_P)\phi_0/(2\pi)$ and can be expressed as

$$V_{out}(t) = \frac{I_P}{2\pi C_P} t \cdot \phi_0 u(t). \tag{15.37}$$

The impulse response is therefore given by

$$h(t) = \frac{I_P}{2\pi C_P} u(t), \tag{15.38}$$

yielding the transfer function

$$\frac{V_{out}}{\Delta\phi}(s) = \frac{I_P}{2\pi C_P} \cdot \frac{1}{s}. \tag{15.39}$$

Consequently, the PFD/CP/LPF combination contains a pole at the origin, a point of contrast to the PD/LPF circuit used in the type I PLL. In analogy with the expression K_{VCO}/s, we call $I_P/(2\pi C_P)$ the "gain" of the PFD and denote it by K_{PFD}.

Example 15.10

Suppose the output quantity of interest in the circuit of Fig. 15.28 is the current injected by the charge pump into the capacitor. Determine the transfer function from $\Delta\phi$ to this current, I_{out}.

[10] We neglect the effect of the narrow reset pulses that appear in the other output.

Solution

Since $V_{out}(s) = I_{out}/(C_P s)$, we have

$$\frac{I_{out}}{\Delta\phi}(s) = \frac{I_P}{2\pi}. \tag{15.40}$$

Let us now construct a linear model of charge-pump PLLs. Shown in Fig. 15.33, the model gives an open-loop transfer function

$$\frac{\Phi_{out}}{\Phi_{in}}(s)|_{\text{open}} = \frac{I_P}{2\pi C_P}\frac{K_{VCO}}{s^2}. \tag{15.41}$$

Since the loop gain has two poles at the origin, this topology is called a "type II" PLL. The closed-loop transfer function, denoted by $H(s)$ for the sake of brevity, is thus equal to

$$H(s) = \frac{\dfrac{I_P K_{VCO}}{2\pi C_P}}{s^2 + \dfrac{I_P K_{VCO}}{2\pi C_P}}. \tag{15.42}$$

This result is alarming because the closed-loop system contains two imaginary poles at $s_{1,2} = \pm j\sqrt{I_P K_{VCO}/(2\pi C_P)}$ and is therefore unstable. The instability arises because the loop gain has only two poles at the origin, (i.e., two ideal integrators). As shown in Fig. 15.34(a), each integrator contributes a constant phase shift of 90°, allowing the system to oscillate at the gain crossover frequency.

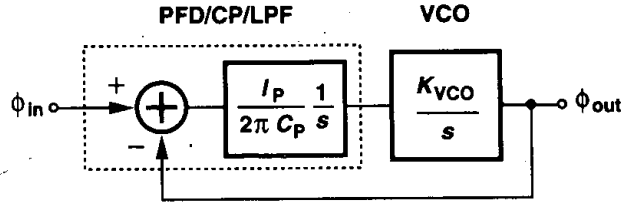


Figure 15.33 Linear model of simple charge-pump PLL.

In order to stabilize the system, we must modify the phase characteristic such that the phase shift is less than 180° at the gain crossover. As shown in Fig. 15.34(b), this is accomplished by introducing a zero in the loop gain, i.e., by adding a resistor in series with the loop filter capacitor (Fig. 15.35). Using the result of Example 15.10, the reader can prove (Problem 15.11) that the PFD/CP/LPF now has a transfer function

$$\frac{V_{out}}{\Delta\phi}(s) = \frac{I_P}{2\pi}\left(R_P + \frac{1}{C_P s}\right). \tag{15.43}$$

It follows that the PLL open-loop transfer function is equal to

$$\frac{\Phi_{out}}{\Phi_{in}}(s)|_{\text{open}} = \frac{I_P}{2\pi}\left(R_P + \frac{1}{C_P s}\right)\frac{K_{VCO}}{s}, \tag{15.44}$$

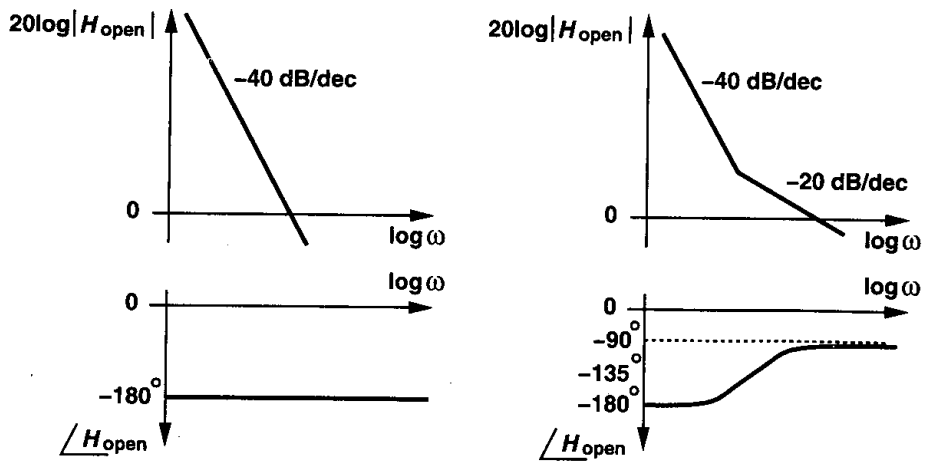


Figure 15.34 (a) Loop gain characteristics of simple charge-pump PLL, (b) addition of zero.

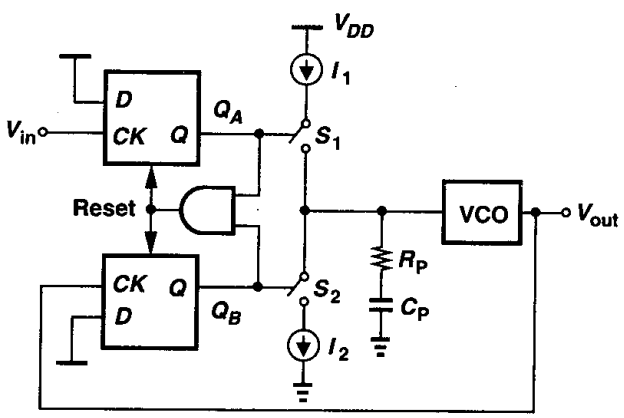


Figure 15.35 Addition of zero to charge-pump PLL.

and hence

$$H(s) = \frac{\dfrac{I_P K_{VCO}}{2\pi C_P}(R_P C_P s + 1)}{s^2 + \dfrac{I_P}{2\pi} K_{VCO} R_P s + \dfrac{I_P}{2\pi C_P} K_{VCO}}. \tag{15.45}$$

The closed-loop system contains a zero at $s_z = -1/(R_P C_P)$. Using the same notation as that for the type I PLL, we have

$$\omega_n = \sqrt{\frac{I_P K_{VCO}}{2\pi C_P}} \tag{15.46}$$

$$\zeta = \frac{R_P}{2}\sqrt{\frac{I_P C_P K_{VCO}}{2\pi}}. \tag{15.47}$$

As expected, if $R_P = 0$, then $\zeta = 0$. With complex poles, the decay time constant is given by $1/(\zeta\omega_n) = 4\pi/(R_P I_P K_{VCO})$.

Stability Issues The stability behavior of type II PLLs is quite different from that of type I PLLs. We begin the analysis with the Bode plots of the loop gain [Eq. (15.44)]. Shown in Fig. 15.36, these plots suggest that if $I_P K_{VCO}$ decreases, the gain crossover frequency

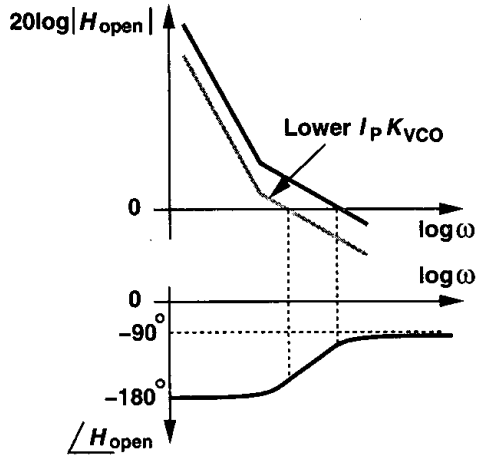


Figure 15.36 Stability degradation of charge-pump PLL as $I_P K_{VCO}$ decreases.

moves toward the origin, *degrading* the phase margin. Predicted by (15.47), this trend is in sharp contrast to that expressed by (15.18) and illustrated in Fig. 15.19.

It is also possible to construct the root locus of the closed-loop system in the complex plane. For $I_P K_{VCO} = 0$ (e.g., $I_P = 0$), the loop is open and both poles lie at the origin. For $I_P K_{VCO} > 0$, we have, $s_{1,2} = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}$, and, since $\zeta \propto \sqrt{I_P K_{VCO}}$, the poles are complex if $I_P K_{VCO}$ is small. The reader can prove (Problem 15.14) that as $I_P K_{VCO}$ increases, s_1 and s_2 move on a circle centered at $\sigma = -1/(R_P C_P)$ with a radius $1/(R_P C_P)$ (Fig. 15.37). The poles return to the real axis at $\zeta = 1$, assuming a value of $-2/(R_P C_P)$. For $\zeta > 1$, the poles remain real, one approaching $-1/(R_P C_P)$ and the other going to $-\infty$ as $I_P K_{VCO} \to +\infty$. Since for complex s_1 and s_2, $\zeta = \cos\varphi$, we observe that as $I_P K_{VCO}$ exceeds zero, the system becomes more stable.

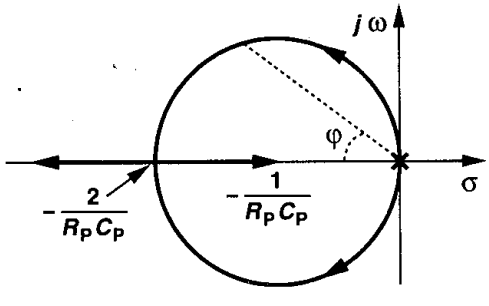


Figure 15.37 Root locus of type II PLL.

The compensated type II PLL of Fig. 15.35 suffers from a critical drawback. Since the charge pump drives the series combination of R_P and C_P, each time a current is injected into the loop filter, the control voltage experiences a large jump. Even in the locked condition, the mismatches between I_1 and I_2 and the charge injection and clock feedthrough of S_1 and S_2 introduce voltage jumps in V_{cont}. The resulting ripple severely disturbs the VCO, corrupting the output phase. To relax this issue, a second capacitor is usually added in parallel with R_P and C_P (Fig. 15.38), suppressing the initial step. The loop filter now is of

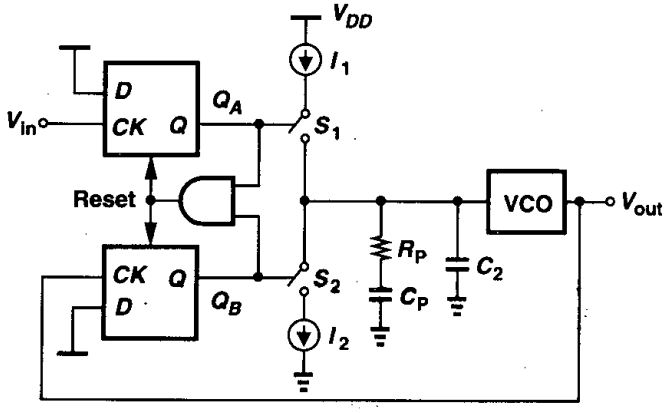


Figure 15.38 Addition of C_2 to reduce ripple on the control line.

second order, yielding a third-order PLL and creating stability difficulties [4]. Nonetheless, if C_2 is about one-fifth to one-tenth of C_P, the closed-loop time and frequency responses remain relatively unchanged.

Equation (15.47) implies that the loop becomes more stable as R_P increases. In reality, as R_P becomes very large, the stability degrades again. This effect is not predicted by the foregoing derivations because we have approximated the discrete-time system by a continuous-time loop. A more accurate analysis is given in [2], but simulations are often necessary to determine the stability bounds of CPPLLs.

15.3 Nonideal Effects in PLLs

15.3.1 PFD/CP Nonidealities

Several imperfections in the PFD/CP circuit lead to high ripple on the control voltage even when the loop is locked. As mentioned earlier, the ripple modulates the VCO frequency, producing a waveform that is no longer periodic. In this section, we study these nonidealities.

The PFD implementation of Fig. 15.24(a) generates narrow, coincident pulses on both Q_A and Q_B even when the input phase difference is zero. As illustrated in Fig. 15.39, if A and B rise simultaneously, so do Q_A and Q_B, thereby activating the reset. That is, even when the PLL is locked, Q_A and Q_B simultaneously turn on the charge pump for a finite period $T_P \approx 10T_D$, where T_D denotes the gate delay (Example 15.8).

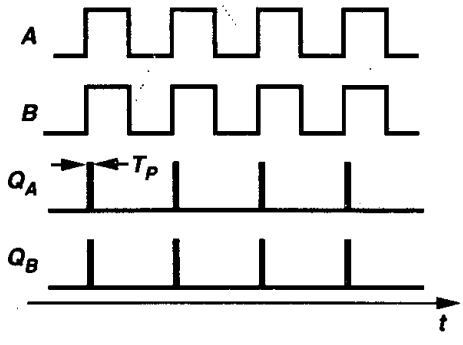


Figure 15.39 Coincident pulses generated by PFD with zero phase difference.

What are the consequences of the reset pulses on Q_A and Q_B? To understand why these pulses are *desirable*, we consider a hypothetical PFD that produces no pulses for a zero input phase difference [Fig. 15.40(a)]. How does such a PFD respond to a small phase error? As shown in Fig. 15.40(b), the circuit generates very narrow pulses on Q_A or Q_B.

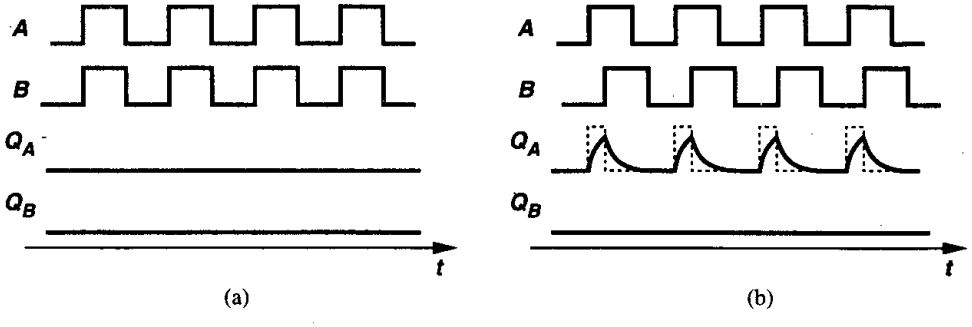


Figure 15.40 Output waveforms of a hypothetical PD with (a) zero input phase difference, and (b) a small input phase difference.

However, owing to the finite risetime and falltime resulting from the capacitance seen at these nodes, the pulse may not find enough time to reach a logical high level, failing to turn on the charge pump switches. In other words, if the input phase difference, $\Delta\phi$, falls below a certain value ϕ_0, then the output voltage of the PFD/CP/LPF combination is no longer a function of $\Delta\phi$. Since, as depicted in Fig. 15.41, for $|\Delta\phi| < \phi_0$ the charge pump injects

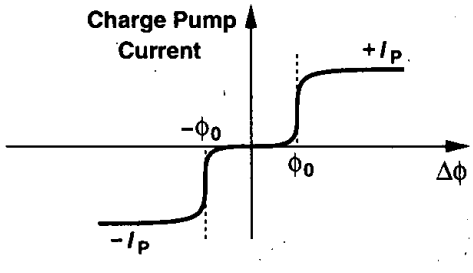


Figure 15.41 Dead zone in the charge pump current.

no current, Eq. (15.41) implies that the loop gain drops to zero and the output phase is not locked. We say the PFD/CP circuit suffers from a dead zone equal to $\pm\phi_0$ around $\Delta\phi = 0$.

The dead zone is highly undesirable because it allows the VCO to accumulate as much random phase error as ϕ_0 with respect to the input while receiving no corrective feedback. Thus, as illustrated in Fig. 15.42, the zero crossing points of the VCO output experience substantial random variations, an effect called "jitter."

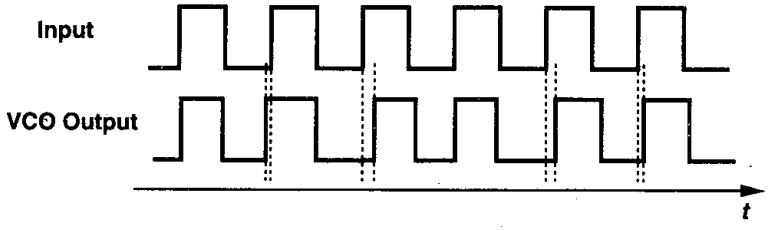


Figure 15.42 Jitter resulting from the dead zone.

Interestingly, the coincident pulses on Q_A and Q_B can eliminate the dead zone. This is because, for $\Delta\phi = 0$, the pulses always turn on the charge pump if they are sufficiently wide. Consequently, as shown in Fig. 15.43, an infinitesimal increment in the phase difference

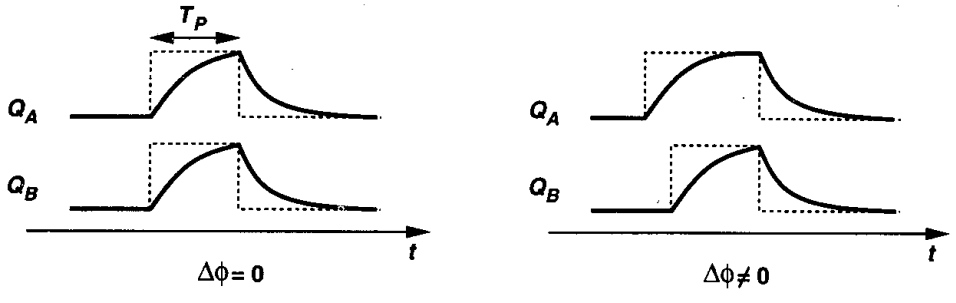


Figure 15.43 Response of actual PD to a small input phase difference.

results in a proportional increase in the net current produced by the charge pump. In other words, the dead zone vanishes if T_P is long enough to allow Q_A and Q_B to reach a valid logical level and turn on the switches in the charge pump.

While eliminating the dead zone, the reset pulses on Q_A and Q_B introduce other difficulties. Let us first implement the charge pump using MOS transistors [Fig. 15.44(a)]. Here, M_1 and M_2 operate as current sources and M_3 and M_4 as switches. The output Q_A is inverted so that when it goes high, M_4 turns on.

The first issue in the circuit of Fig. 15.44(a) stems from the delay difference between $\overline{Q_A}$ and Q_B in turning on their respective switches. As shown in Fig. 15.44(b), the net current injected by the charge pump into the loop filter jumps to $+I_P$ and $-I_P$, disturbing the oscillator control voltage periodically even if the loop is locked. To suppress this effect,

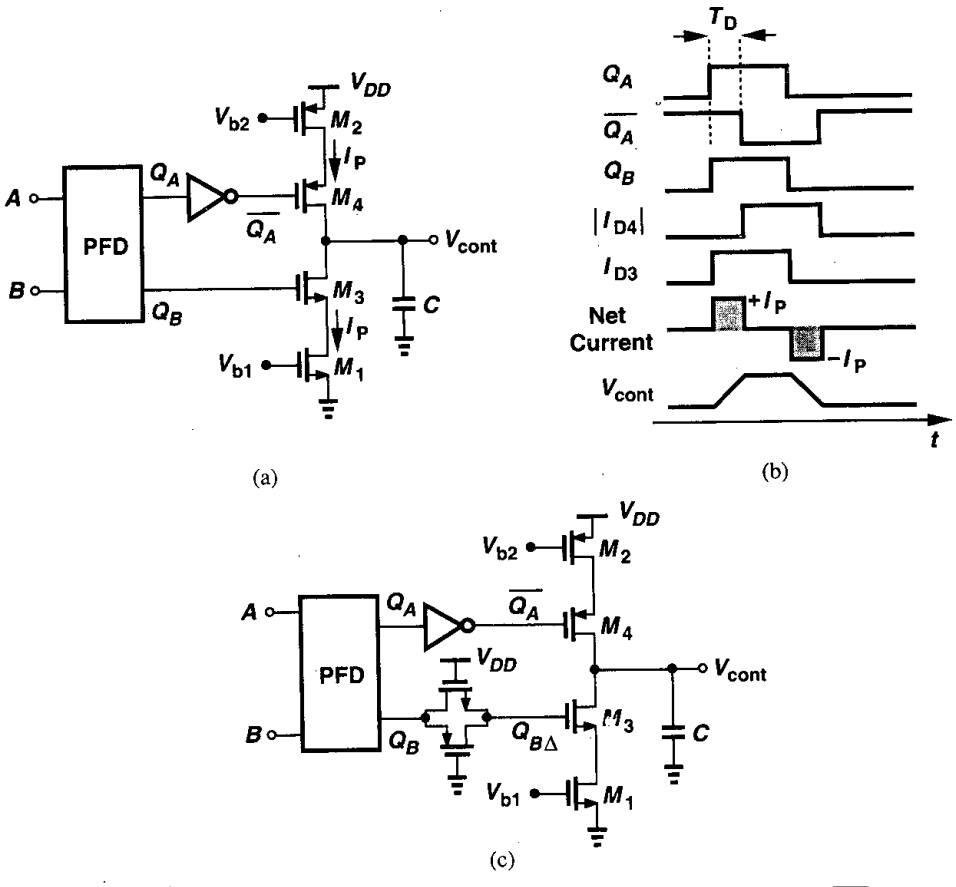


Figure 15.44 (a) Implementation of charge pump, (b) effect of skew between $\overline{Q_A}$ and Q_B, (c) suppression of skew by a pass gate.

a complementary pass gate can be interposed between Q_B and the gate of M_3, equalizing the delays [Fig. 15.44(c)].

The second issue in the CP of Fig. 15.44(c) relates to the mismatch between the drain currents of M_1 and M_2. As depicted in Fig. 15.45(a), even with perfect alignment of the UP and DOWN pulses, the net current produced by the charge pump is nonzero, changing V_{cont} by a constant increment at each phase comparison instant. How does the PLL respond to this error? For the loop to remain locked, the average value of the control voltage must remain constant. The PLL therefore creates a phase error between the input and the output such that the net current injected by the CP in every cycle is zero [Fig. 15.45(b)]. The relationship between the current mismatch and the phase error is determined in Problem 15.12. It is important to note that (1) the control voltage still experiences a periodic ripple, (2) owing to the low output impedance of short-channel MOSFETs, the current mismatch *varies* with the output voltage (i.e., with the VCO frequency), and (3) the clock feedthrough and

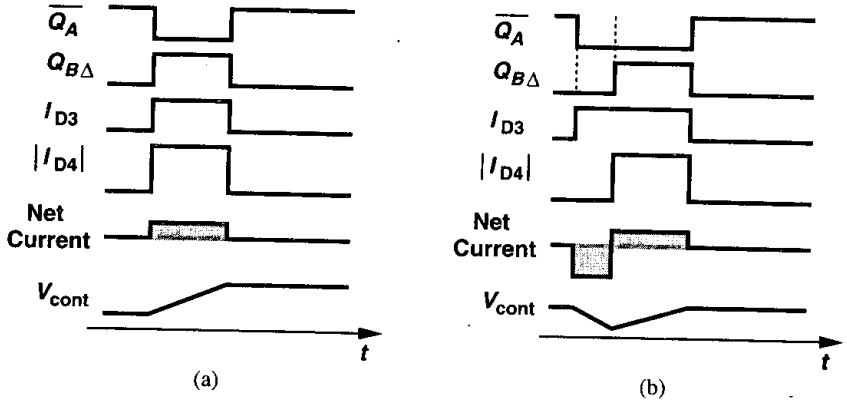


Figure 15.45 Effect of UP and DOWN current mismatch.

charge injection mismatch between M_3 and M_4 further increases both the phase error and the ripple.

The third issue in the circuit of Fig. 15.44(c) originates from the finite capacitance seen at the drains of the current sources. Suppose, as illustrated in Fig. 15.46(a), S_1 and S_2 are off, allowing M_1 to discharge X to ground and M_2 to charge Y to V_{DD}. At the next phase comparison instant, both S_1 and S_2 turn on, V_X rises, V_Y falls, and $V_X \approx V_Y \approx V_{cont}$ if the voltage drop across S_1 and S_2 is neglected [Fig. 15.46(b)]. If the phase error is zero and

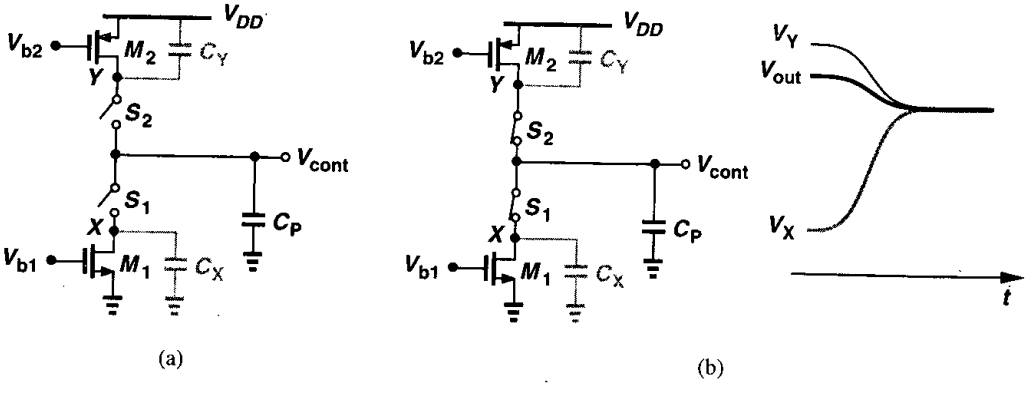


Figure 15.46 Charge sharing between C_P and capacitances at X and Y.

$I_{D1} = |I_{D2}|$, does V_{cont} remain constant after the switches turn on? Even if $C_X = C_Y$, the change in V_X is not equal to that in V_Y. For example, if V_{cont} is relatively high, V_X changes by a large amount and V_Y by a small amount. The difference between the two changes must therefore be supplied by C_P, leading to a jump in V_{cont}.

The above charge sharing phenomenon can be suppressed by "bootstrapping." Illustrated in Fig. 15.47 [3], the idea is to "pin" V_X and V_Y to V_{cont} after phase comparison is finished. When S_1 and S_2 turn off, S_3 and S_4 turn on, allowing the unity-gain amplifier to hold nodes

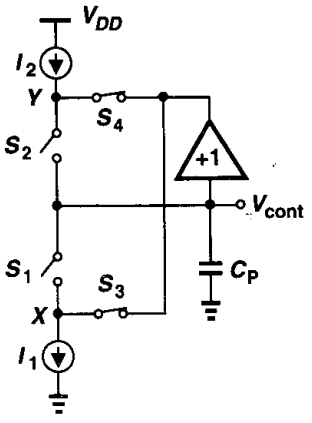


Figure 15.47 Bootstrapping X and Y to minimize charge sharing.

X and Y at a potential equal to V_{cont}. Note that the amplifier need not provide much current because $I_1 \approx I_2$. At the next phase comparison instant, S_1 and S_2 turn on, S_3 and S_4 turn off, and V_X and V_Y begin with a value equal to V_{cont}. Thus, no charge sharing occurs between C_P and the capacitances at X and Y.

15.3.2 Jitter in PLLs

The response of phase-locked loops to jitter is of extreme importance in most applications. We first describe the concepts of jitter and the rate of change of jitter.

As shown in Fig. 15.48, a strictly periodic waveform, $x_1(t)$, contains zero crossings that are evenly spaced in time. Now consider the nearly periodic signal $x_2(t)$, whose period

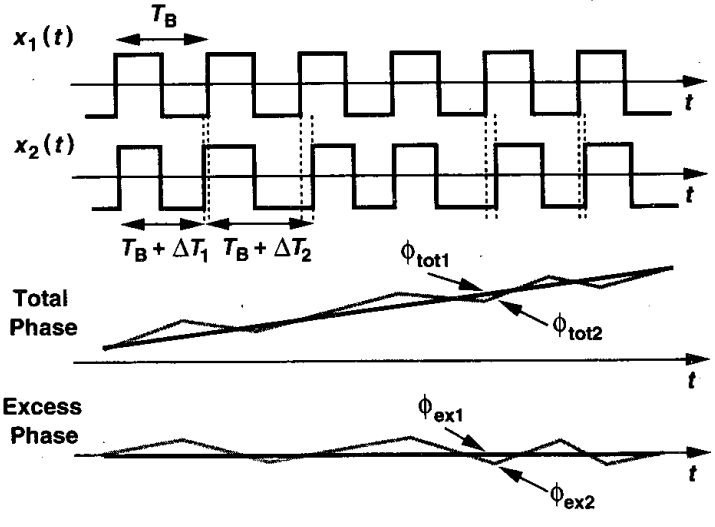


Figure 15.48 Ideal and jittery waveforms.

experiences small changes, deviating the zero crossings from their ideal points. We say the latter waveform suffers from jitter.[11] Plotting the total phase, ϕ_{tot}, and the excess phase, ϕ_{ex}, of the two waveforms, we observe that jitter manifests itself as variation of the excess phase with time. In fact, ignoring the harmonics above the fundamental, we can write $x_1(t) = A\cos\omega t$ and $x_2(t) = A\cos[\omega t + \phi_n(t)]$, where $\phi_n(t)$ models the variation of the period.[12]

The rate at which the jitter varies is also important. Consider the two jittery waveforms depicted in Fig. 15.49. The first signal, $y_1(t)$, experiences "slow jitter" because its instantaneous frequency varies slowly from one period to the next. The second signal, $y_2(t)$,

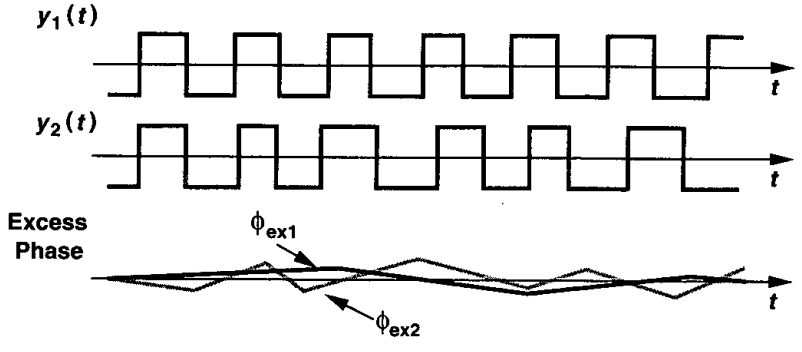


Figure 15.49 Illustration of slow and fast jitter.

experiences "fast jitter." The rate of change is also evident from the excess phase plots of the two waveforms.

Two jitter phenomena in phase-locked loops are of great interest: (a) the input exhibits jitter, and (b) the VCO produces jitter. Let us study each case, assuming the input and output waveforms are expressed as $x_{in}(t) = A\cos[\omega t + \phi_{in}(t)]$ and $x_{out}(t) = A\cos[\omega t + \phi_{out}(t)]$.

The transfer functions derived for type I and type II PLLs have a low-pass characteristic, suggesting that if $\phi_{in}(t)$ varies rapidly, then $\phi_{out}(t)$ does not fully track the variations. In other words, slow jitter at the input propagates to the output unattenuated but fast jitter does not. We say the PLL low-pass filters $\phi_{in}(t)$.

Now suppose the input is strictly periodic but the VCO suffers from jitter. Viewing jitter as random phase variations, we construct the model depicted in Fig. 15.50, where the input excess phase is set to zero [i.e., $x_{in}(t) = A\cos\omega t$] and a random component Φ_{VCO} is added to the output of the VCO to represent its jitter. The reader can show that the transfer function from Φ_{VCO} to Φ_{out} for a type II PLL is equal to

$$\frac{\Phi_{out}}{\Phi_{VCO}}(s) = \frac{s^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \tag{15.48}$$

[11] Jitter is quantified by several different mathematical definitions, e.g., as in [5].

[12] The quantity $\phi_n(t)$ (or more commonly its spectrum) is called the "phase noise." In this book, we assume the jitter is uniquely represented by $\phi_n(t)$.

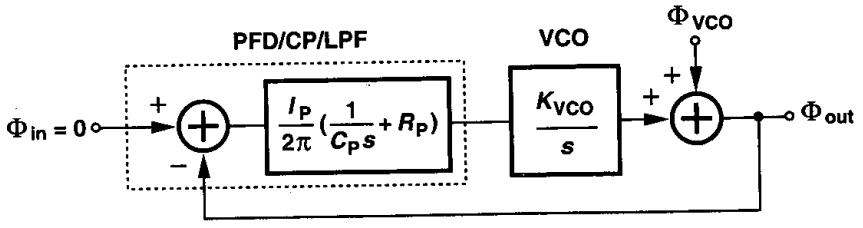


Figure 15.50 Effect of VCO jitter.

Interestingly, the characteristic has a high-pass nature, indicating that slow jitter components generated by the VCO are suppressed but fast jitter components are not. This can be understood with the aid of Fig. 15.50: If $\phi_{VCO}(t)$ changes slowly (e.g., the oscillation period drifts with temperature), then the comparison with $\phi_{in} = 0$ (i.e., a perfectly periodic signal) generates a slowly varying error that propagates through the LPF and adjusts the VCO frequency, thereby counteracting the change in ϕ_{VCO}. On the other hand, if ϕ_{VCO} varies rapidly, (e.g., high-frequency noise modulates the oscillation period), then the error produced by the phase detector is heavily attenuated by the poles in the loop, failing to correct for the change.

Figure 15.51 conceptually summarizes the response of PLLs to input jitter and VCO jitter. Depending on the application and the environment, one or both sources may be significant, requiring an optimum choice of the loop bandwidth.

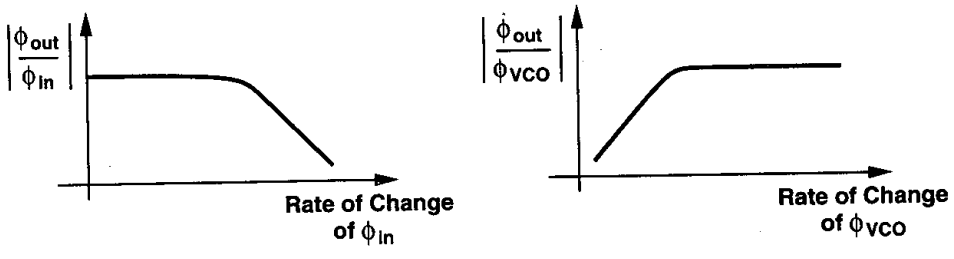


Figure 15.51 Transfer functions of jitter from input and VCO to the output.

15.4 Delay-Locked Loops

A variant of PLLs that has become popular in the past ten years is the delay-locked loop. To arrive at the concept, let us begin with an example. Suppose an application requires four clock phases with a precise spacing of $\Delta T = 1$ ns between consecutive edges [Fig. 15.52(a)]. How should these phases be generated? We can use a two-stage differential ring oscillator[13] to produce the four phases, but how do we guarantee that $\Delta T = 1$ ns

[13] As explained in Chapter 14, a simple two-stage CMOS ring oscillator may not oscillate. This example is merely for illustration purposes.

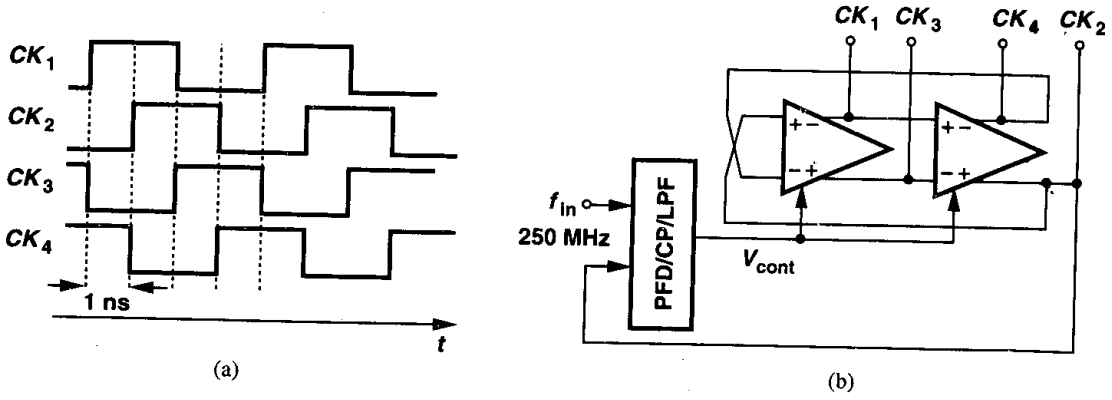


Figure 15.52 (a) Clock phases with edge-to-edge delay of 1 ns, (b) use of a phase-locked ring oscillator to generate the clock phases.

despite process and temperature variations? This requires that the oscillator be locked to a 250-MHz reference so that the output period is exactly equal to 4 ns [Fig. 15.52(b)].

An alternative approach to generating the clock phases of Fig. 15.52(a) is to apply the input clock to four delay stages in a cascade. Illustrated in Fig. 15.53(a), this technique nonetheless does not produce a well-defined edge spacing because the delay of each stage

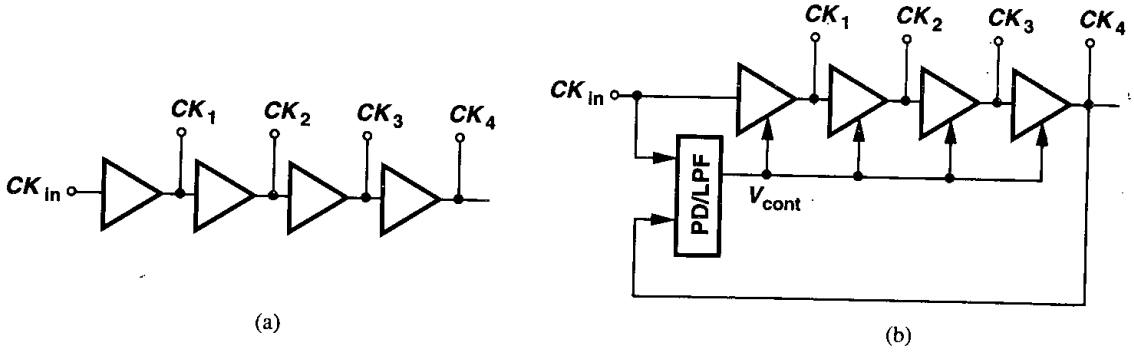


Figure 15.53 (a) Generation of clock edges by delay stages, (b) simple delay-locked loop.

varies with process and temperature. Now consider the circuit shown in Fig. 15.53(b), where the phase difference between CK_{in} and CK_4 is sensed by a phase detector, a proportional average voltage, V_{cont}, is generated, and the delay of the stages is adjusted with negative feedback. For a large loop gain, the phase difference between CK_{in} and CK_4 is small, that is, the four stages delay the clock by almost exactly one period, thereby establishing precise edge spacing.[14] This topology is called a delay-locked loop to emphasize that it incorporates a voltage-controlled delay line (VCDL) rather than a VCO. In practice, a charge pump is

[14]The total delay through the four stages may be equal to two or more periods. We return to this issue later.

interposed between the PD and the LPF to achieve an infinite loop gain. Each delay stage may be based on one of the ring oscillator stages described in Chapter 14.

The reader may wonder about the advantages of DLLs over PLLs. First, delay lines are generally less susceptible to noise than oscillators are because corrupted zero crossings of a waveform disappear at the end of a delay line whereas they are recirculated in an oscillator, thereby experiencing more corruption. Second, in the VCDL of Fig. 15.53(b), a change in the control voltage immediately changes the delay, that is, the transfer function $\Phi_{out}(s)/V_{cont}(s)$ is simply equal to the gain of the VCDL, K_{VCDL}. Thus, the feedback system of Fig. 15.53(b) has the same order as the LPF and its stability and settling issues are more relaxed than those of a PLL.

Example 15.11

Determine the closed-loop transfer function of the DLL shown in Fig. 15.54.

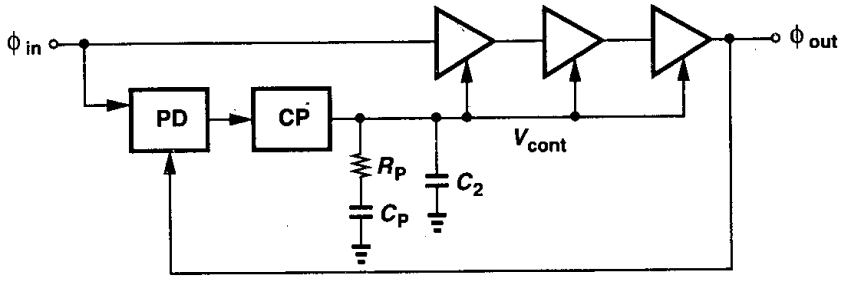


Figure 15.54

Solution

From Example 15.10, we write the transfer function of the PD/CP/LPF combination as

$$\frac{V_{cont}}{\Delta\Phi}(s) = \frac{I_P}{2\pi}\left[\left(R_P + \frac{1}{C_P s}\right)\Big\|\frac{1}{C_2 s}\right] \tag{15.49}$$

$$= \frac{I_P}{2\pi}\frac{R_P C_P s + 1}{(R_P C_P C_2 s + C_P + C_2)s}. \tag{15.50}$$

The closed-loop transfer function is thus equal to

$$\frac{\Phi_{out}}{\Phi_{in}}(s)|_{\text{closed}} = \frac{\frac{I_P K_{VCDL}}{2\pi}(R_P C_P s + 1)}{R_P C_P C_2 s^2 + [C_P + C_2 + I_P K_{VCDL} R_P C_P/(2\pi)]s + I_P K_{VCDL}/(2\pi)}. \tag{15.51}$$

This transfer function can be used to determine how ϕ_{out} settles if ϕ_{in} experiences a change. Note that in practice R_P may not be needed because the loop contains only one pole at the origin.

The principal drawback of DLLs is that they cannot generate a variable output frequency. This issue becomes clearer when we study the frequency synthesis capabilities of PLLs in Section 15.5.1. DLLs may also suffer from locked delay ambiguity. That is, if the total delay of the four stages in Fig. 15.53(b) can vary from below T_{in} to above $2T_{in}$, then the loop may lock with a CK_{in}-to-CK_4 delay equal to either T_{in} or $2T_{in}$. This ambiguity proves detrimental if the DLL must provide precisely-spaced clock edges because the edge-to-edge delay may settle to $2T_{in}/4$ rather than $T_{in}/4$. In such cases, additional circuitry is necessary to avoid the ambiguity. Also, mismatches between the delay stages and their load capacitances introduce error in the edge spacing, requiring large devices and careful layout.

15.5 Applications

After nearly 70 years since its invention, phase locking continues to find new applications in electronics, communication, and instrumentation. Examples include memories, microprocessors, hard disk drive electronics, RF and wireless transceivers, and optical fiber receivers.

The reader may recall from Section 15.1.2 that a PLL appears no more useful than a short piece of wire because both guarantee a small phase difference between the input and the output. In this section, we present a number of applications that demonstrate the versatility of phase locking. The concepts described below have been the topic of numerous books and papers, e.g., [6, 7].

15.5.1 Frequency Multiplication and Synthesis

Frequency Multiplication A PLL can be modified such that it multiplies its input frequency by a factor of M. To arrive at the implementation, we exploit an analogy with voltage multiplication. As depicted in Fig. 15.55(a), a feedback system amplifies the input

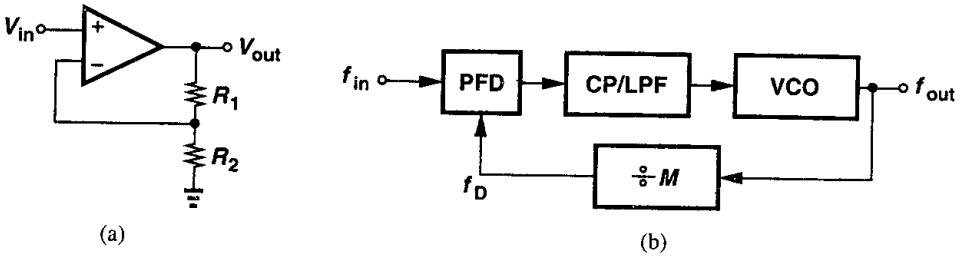


Figure 15.55 (a) Voltage amplification and (b) frequency multiplication.

voltage by a factor of M if the output voltage is divided by M [i.e., if $R_2/(R_1 + R_2) = 1/M$] and the result is compared with the input. Thus, as shown in Fig. 15.55(b), if the output *frequency* of a PLL is divided by M and applied to the phase detector, we have $f_{out} = Mf_{in}$. From another point of view, since $f_D = f_{out}/M$ and f_D and f_{in} must be equal in the locked condition, the PLL multiplies f_{in} by M. The $\div M$ circuit is realized as a counter that produces one output pulse for every M input pulses.

As with voltage division in Fig. 15.55(a), the feedback divider in the loop of Fig. 15.55(b) alters the system characteristics. Using (15.44), we rewrite (15.45) as

$$H(s) = \frac{\frac{I_P}{2\pi}\left(R_P + \frac{1}{C_P s}\right)\frac{K_{VCO}}{s}}{1 + \frac{1}{M}\frac{I_P}{2\pi}\left(R_P + \frac{1}{C_P s}\right)\frac{K_{VCO}}{s}} \tag{15.52}$$

$$= \frac{\frac{I_P K_{VCO}}{2\pi C_P}(R_P C_P s + 1)}{s^2 + \frac{I_P}{2\pi}\frac{K_{VCO}}{M}R_P s + \frac{I_P}{2\pi C_P}\frac{K_{VCO}}{M}}. \tag{15.53}$$

Note that $H(s) \to M$ as $s \to 0$, i.e., phase or frequency changes at the input result in an M-fold change in the corresponding output quantity. Comparing the denominators of (15.45) and (15.53), we observe that frequency division in the loop manifests itself as division of K_{VCO} by M. In other words, as far as the poles of the closed-loop system are concerned, we can assume the oscillator and the divider form a VCO with an equivalent gain of K_{VCO}/M. This is of course to be expected because, for the VCO/divider cascade shown Fig. 15.56, we have

$$\omega_{out} = \frac{\omega_0 + K_{VCO}V_{cont}}{M} \tag{15.54}$$

$$= \frac{\omega_0}{M} + \frac{K_{VCO}}{M}V_{cont}. \tag{15.55}$$

Thus, the combination cannot be distinguished from a VCO having an intercept frequency of ω_0/M and a gain of K_{VCO}/M.

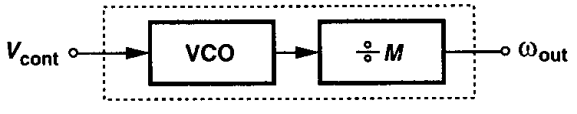


Figure 15.56 Equivalency of VCO/divider combination to a single VCO.

The foregoing discussion suggests that (15.46) and (15.47) can be respectively rewritten as

$$\omega_n = \sqrt{\frac{I_P}{2\pi C_P}\frac{K_{VCO}}{M}} \tag{15.56}$$

$$\zeta = \frac{R_P}{2}\sqrt{\frac{I_P C_P}{2\pi}\frac{K_{VCO}}{M}}. \tag{15.57}$$

Also, the decay time constant is modified to $(\zeta\omega_n)^{-1} = 4\pi M/(R_P I_P K_{VCO})$. It follows that inserting a divider in a type II loop degrades both the stability and the settling speed, requiring a proportional increase in the charge pump current.

The frequency-multiplying loop of Fig. 15.55(b) exhibits two interesting properties. First, unlike the voltage amplifier of Fig. 15.55(a), the PLL provides a multiplication factor *exactly* equal to M, an attribute resulting from the infinite loop gain and expressed by Eq. (15.53). Second, the output frequency can be varied by changing the divide ratio M, an extremely useful property in synthesizing frequencies. Note that DLLs cannot perform such synthesis.

Frequency Synthesis Some systems require a periodic waveform whose frequency (a) must be very accurate (e.g., exhibit an error less than 10 ppm), and (b) can be varied in very fine steps (e.g., in steps of 30 kHz from 900 MHz to 925 MHz). Commonly encountered in wireless transceivers, such requirements can be met through frequency multiplication by PLLs.

Figure 15.57 shows the architecture of a phase-locked frequency synthesizer. The channel control input is a digital word that varies the value of M. Since $f_{out} = Mf_{REF}$, the relative accuracy of f_{out} is equal to that of f_{REF}. For this reason, f_{REF} is derived from a stable, low-noise crystal oscillator. Note that f_{out} varies in steps equal to f_{REF} if M changes by one each time.

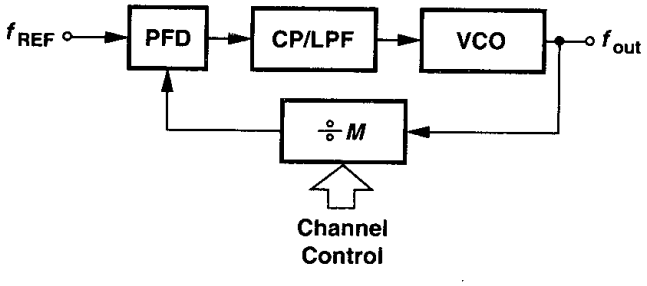


Figure 15.57 Frequency synthesizer.

CMOS frequency synthesizers achieving gigahertz output frequencies have been reported. Issues such as noise, sidebands, settling speed, frequency range, and power dissipation continue to challenge synthesizer designers.

15.5.2 Skew Reduction

The earliest usage of phase locking in digital systems was for skew reduction. Suppose a synchronous pair of data and clock lines enter a large digital chip as shown in Fig. 15.58. Since the clock typically drives a large number of transistors and long interconnects, it is first applied to a large buffer. Thus, the clock distributed on the chip may suffer from substantial skew with respect to the data, an undesirable effect because it reduces the timing budget for on-chip operations.

Now consider the circuit shown in Fig. 15.59, where CK_{in} is applied to an on-chip PLL and the buffer is placed *inside* the loop. Since the PLL guarantees a nominally-zero phase difference between CK_{in} and CK_B, the skew is eliminated. From another point of view, the constant phase shift introduced by the buffer is divided by the infinite loop gain of

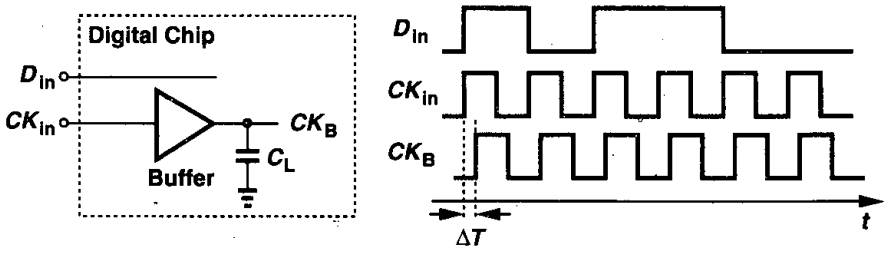


Figure 15.58 Skew between data and buffered clock.

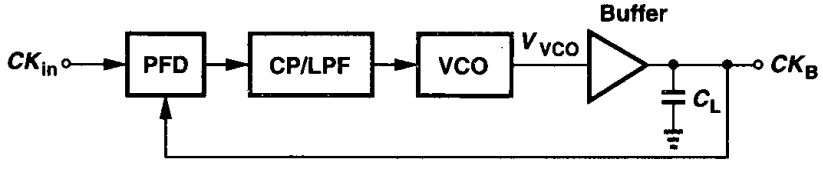


Figure 15.59 Use of a PLL to eliminate skew.

the feedback system. Note that the VCO output, V_{VCO}, may not be aligned with CK_{in}, a nonetheless unimportant issue because V_{VCO} is not used.

Example 15.12

Construct the voltage-domain counterpart of the loop shown in Fig. 15.59.

Solution

The buffer creates a constant phase shift in the signal generated by the VCO. The voltage-domain counterpart therefore assumes the topology shown in Fig. 15.60. We have

$$(V_{in} - V_{out})A + V_M = V_{out} \tag{15.58}$$

and hence

$$V_{out} = \frac{AV_{in} + V_M}{1 + A}. \tag{15.59}$$

As $A \to \infty$, $V_{out} \to V_{in}$.

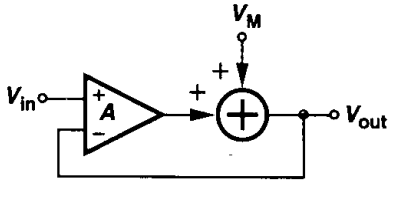


Figure 15.60

We should note that the skew can be suppressed by a delay-locked loop as well. In fact, if frequency multiplication is not required, DLLs are preferred because they are less susceptible to noise.

15.5.3 Jitter Reduction

Recall from Section 15.3.2 that PLLs suppress fast jitter components at the input. For example, if a 1-GHz jittery signal is applied to a PLL having a bandwidth of 10 MHz, then input jitter components that vary faster than 10 MHz are attenuated. In a sense, the phase-locked loop operates as a narrowband filter centered around 1 GHz with a total bandwidth of 20 MHz. This is another important and useful property of PLLs.

Many applications must deal with jittery waveforms. Random binary signals experience jitter because of (a) crosstalk on the chip and in the package (Chapter 18), (b) package parasitics (Chapter 18), (c) additive electronic noise of devices, etc. Such waveforms are typically "retimed" by a low-noise clock so as to reduce the jitter. Illustrated in Fig. 15.61(a), the idea is to resample the midpoint of each bit by a D flipflop that

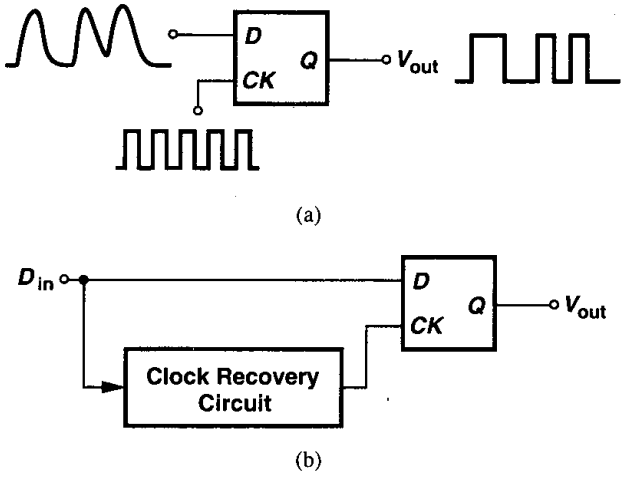


Figure 15.61 (a) Retiming data with D flipflop driven by a low-noise clock, (b) use of a phase-locked clock recovery circuit to generate the clock.

is driven by the clock. However, in many applications, the clock may not be available independently. For example, an optical fiber carries only the random data stream, providing no separate clock waveform at the receive end. The circuit of Fig. 15.61(a) is therefore modified as shown in Fig. 15.61(b), where a "clock recovery circuit" (CRC) produces the clock from the data. Employing phase locking with a relatively narrow loop bandwidth, the CRC minimizes the effect of the input jitter on the recovered clock.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume $V_{DD} = 3$ V where necessary. Also, assume all transistors are in saturation.

15.1. The Gilbert cell (Chapter 4) operates as an XOR gate with large input swings and as an analog multiplier with small input swings. Prove that an analog multiplier can be used to detect the phase difference between two sinusoids. Is the input-output characteristic of such a phase detector linear?

15.2. Redraw the waveforms of Fig. 15.4(b) if the VCO frequency is lowered at $t = t_1$. If the phase error between V_{CK} and V_{VCO} before $t = t_1$ is equal to ϕ_0 and f_{VCO} is lowered from f_H to f_L, determine the minimum $t_2 - t_1$ that is sufficient for phase alignment.

15.3. Explain why the low-pass filter in Fig. 15.5(b) cannot be replaced by a high-pass filter.

15.4. A PLL using an XOR gate as a phase detector locks with $\phi_{in} - \phi_{out} \approx 90°$ if $K_{PD}K_{VCO}$ is large. Explain why?

15.5. Using the characteristic of Fig. 15.3 as an example, explain why the polarity of feedback in a PLL (without frequency detection) is unimportant. (Hint: prove that the loop locks regardless of whether the initial phase difference falls in the positive-slope region or the negative-slope region.)

15.6. Assuming a first-order LPF in Fig. 15.14, determine the transfer function Φ_{out}/Φ_{ex}, where Φ_{out} denotes the excess phase of V_{out}.

15.7. A VCO used in a type I PLL exhibits nonlinearity in its input-output characteristic, i.e., K_{VCO} varies across the tuning range. If the damping ratio must remain between 1 and 1.5, how much variation can be tolerated in K_{VCO}?

15.8. Prove that in the root locus of Fig. 15.20, $\cos\theta = \zeta$.

15.9. A type I PLL incorporates a VCO with $K_{VCO} = 100$ MHz/V, a PD with $K_{PD} = 1$ V/rad, and an LPF with $\omega_{LPF} = 2\pi(1\text{ MHz})$. Determine the step response of the PLL.

15.10. Explain why in the charge-pump PLL of Fig. 15.35, the control voltage of the VCO cannot be connected to the top plate of C_P.

15.11. Prove that the transfer function of the PFD/CP/LPF circuit in Fig. 15.35 is given by Eq. (15.43).

15.12. As illustrated in Fig. 15.45, mismatches between the UP and DOWN currents translate to phase offset at the input of a CPPLL. With the aid of the waveforms in Fig. 15.45, calculate the phase offset in terms of current mismatch.

15.13. For a VCO, we have $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$. The control line experiences a small sinusoidal ripple, $V_{cont} = V_m \cos\omega_m t$. If the VCO is followed by a $\div M$ circuit, determine the output spectrum of the divider. Consider two cases: $\omega_0/M > \omega_m$ and $\omega_0/M < \omega_m$.

15.14. Prove that the root locus of a type II PLL is as shown in Fig. 15.37.

15.15. Determine the transfer function Φ_{out}/Φ_{ex} for the circuit of Fig. 15.14 if the PLL is modified to the architecture of Fig. 15.35.

15.16. When a charge-pump PLL incorporating a PFD is turned on, the VCO frequency may be far from the input frequency. Explain why the order of the PLL transfer function is lower by one while the PFD operates as a frequency detector.

References

1. R. E. Best, *Phase-Locked Loops,* Second Ed., New York: McGraw-Hill, 1993.
2. F. M. Gardner, *Phaselock Techniques,* Second Ed., New York: Wiley & Sons, 1979.
3. M. G. Johnson and E. L. Hudson, "A Variable Delay Line PLL for CPU-Coprocessor Synchronization," *IEEE Journal of Solid-State Circuits,* vol. 23, pp. 1218–1223, Oct. 1988.
4. F. M. Gardner, "Charge-Pump Phase-Locked Loops," *IEEE Trans. Comm.*, vol. COM-28, pp.1849–1858, Nov. 1980.
5. F. Herzel and B. Razavi, "A Study of Oscillator Jitter Due to Supply and Substrate Noise," *IEEE Transactions on Circuits and Systems, Part II,* vol.46, pp.56–62, Jan. 1999.
6. W. F. Egan, *Frequency Synthesis by Phase Lock,* New York: Wiley & Sons, 1981.
7. J. A. Crawford, *Frequency Synthesizer Design Handbook,* New York: Artech House, 1994.

Chapter 16

Short-Channel Effects and Device Models

The square-law characteristics derived for MOSFETs in Chapter 2 provide moderate accuracies for devices with minimum channel lengths of greater than 4 μm, a value corresponding to technologies in production in the early 1980s. As device dimensions continue to scale down, reaching below 0.2 μm by the year 2000, higher order effects necessitate more complex models so as to attain enough accuracy in simulations.

The problem of device models in CMOS technology has constantly haunted analog designers, manifesting itself as substantial discrepancies between simulated and measured results. A number of comprehensive books [1, 2, 3] and hundreds of papers deal with the subject in great detail, but 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. Knowledge of these issues also proves useful in interpreting the anomalies that the designer may encounter in SPICE simulations.

We first describe the ideal scaling theory of MOS transistors. Next, we study short-channel effects such as threshold voltage variation, velocity saturation, and the dependence of the output impedance on the drain-source voltage. We then review MOS device models, including Levels 1–3 and the BSIM series. Finally, we discuss charge and capacitance modeling, temperature dependence, and process corners.

16.1 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. In a paper published in 1974 [4], Dennard et al. recognized the tremendous potential of scaling MOS transistors, making predictions about speed and power dissipation of digital CMOS circuits as devices are shrunk.

The ideal scaling theory follows three rules: (1) reduce all lateral and vertical dimensions by $\alpha(>1)$; (2) reduce the threshold voltage and the supply voltage by α; (3) increase all of the doping levels by α (Fig. 16.1). Since the dimensions and voltages scale together, all electric fields in the transistor remain constant, hence the name "constant-field scaling." Note that

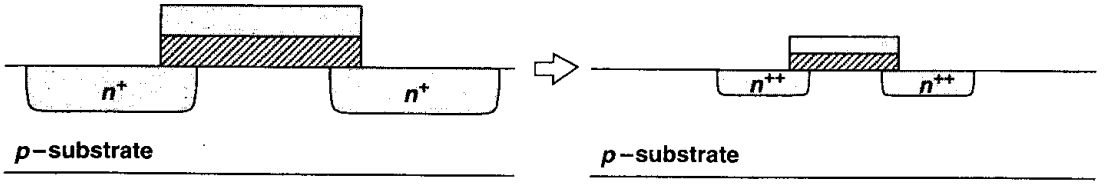


Figure 16.1 Ideal scaling of MOS transistor.

W, L, t_{ox}, V_{DD}, V_{TH}, and the depth and perimeter of the source and drain junctions scale down by α.

Let us examine the saturation drain current of a square-law device after scaling. Writing

$$I_{D,scaled} = \frac{1}{2}\mu_n(\alpha C_{ox})\left(\frac{W/\alpha}{L/\alpha}\right)\left(\frac{V_{GS}}{\alpha} - \frac{V_{TH}}{\alpha}\right)^2 \tag{16.1}$$

$$= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2\frac{1}{\alpha}, \tag{16.2}$$

we observe that the current capability of the transistor *drops* by a factor of α. Note that the same result applies for the drain current in the triode region. The advantage of scaling, however, lies in the reduction of capacitances and power dissipation. The total channel capacitance is

$$C_{ch,scaled} = \frac{W}{\alpha}\frac{L}{\alpha}(\alpha C_{ox}) \tag{16.3}$$

$$= \frac{1}{\alpha}WLC_{ox}. \tag{16.4}$$

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

$$W_d = \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)(\phi_B + V_R)}, \tag{16.5}$$

where N_A and N_D denote the doping levels of the two sides of the junction, $\phi_B = V_T \ln(N_A N_D/n_i^2)$, and V_R is the reverse-bias voltage. The built-in potential, ϕ_B, is a weak function of $N_A N_D$ and in fact it *increases* if $N_A N_D$ is scaled up by α^2. For now, we assume $V_R \gg \phi_B$ so that

$$W_{d,scaled} \approx \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{\alpha N_A} + \frac{1}{\alpha N_D}\right)\frac{V_R}{\alpha}} \tag{16.6}$$

$$\approx \frac{1}{\alpha}\sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)V_R}. \tag{16.7}$$

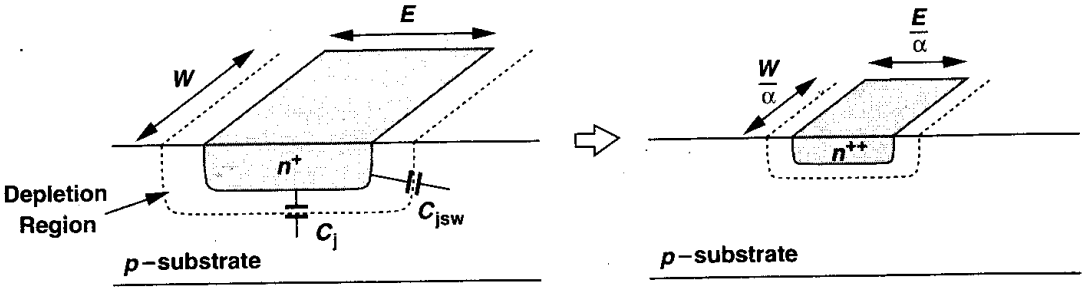


Figure 16.2 Scaling of S/D junction capacitances.

Thus, as with other dimensions, the width of each depletion region scales down by α, increasing the depletion region capacitance per unit area by the same factor.

As illustrated in Fig. 16.2, 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}, on the other hand, remains constant because the depth of the junction is reduced by α. It follows that

$$C_{S/D,scaled} = \frac{W}{\alpha}\frac{E}{\alpha}(\alpha C_j) + 2\left(\frac{W}{\alpha} + \frac{E}{\alpha}\right)(C_{jsw}) \tag{16.8}$$

$$= [WEC_j + 2(W+E)C_{jsw}]\frac{1}{\alpha}. \tag{16.9}$$

All of the capacitances therefore decrease by the scaling factor.

In digital applications, the scaling of the gate delay and power dissipation is of interest. Approximating the delay of a CMOS inverter by $T_d = (C/I)V_{DD}$ (Fig. 16.3), we have

$$T_{d,scaled} = \frac{C/\alpha}{I/\alpha}\frac{V_{DD}}{\alpha} \tag{16.10}$$

$$= \left(\frac{C}{I}V_{DD}\right)\frac{1}{\alpha}. \tag{16.11}$$

We conclude that the speed of digital circuits can potentially increase by the scaling factor. For power dissipation, we write $P = fCV_{DD}^2$, where f is the operating frequency. Thus,

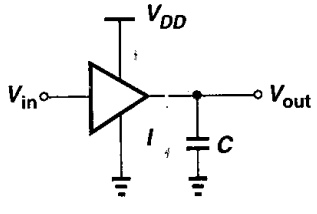


Figure 16.3 CMOS inverter.

$P_{scaled} = f(C/\alpha)(V_{DD}/\alpha)^2 = fCV_{DD}^2/\alpha^3$, if f and the number of gates in the circuit remain constant. Note that the layout density, i.e., the number of transistors per unit area, also scales by α^2.

The reduction of power and delay and the increase in circuit density make scaling extremely attractive for digital systems. Based on these observations, Gordon Moore predicted in 1975 [5] that MOS device dimensions would continue to scale down by a factor of two every three years and the number of transistors per chip would double every one to two years. Such trends have indeed persisted over the past 25 years.

Let us now consider the effect of ideal scaling in analog circuits. Writing the transconductance as

$$g_{m,scaled} = \mu(\alpha C_{ox})\frac{W/\alpha}{L/\alpha}\frac{V_{GS} - V_{TH}}{\alpha} \tag{16.12}$$

$$= \mu C_{ox}\frac{W}{L}(V_{GS} - V_{TH}), \tag{16.13}$$

we note that the transconductance remains constant if all of the dimensions and voltages (and currents) scale down. To calculate the output impedance in saturation, we first observe from Fig. 16.4 and Eq. (16.7) that the width of the depletion region around the drain decreases by α, and hence $\Delta L/L$ remains constant. Since $\lambda = (\Delta L/L)/V_{DS}$ (Chapter 2), λ increases by α and

$$r_{O,scaled} = \frac{1}{\alpha\lambda\dfrac{I_D}{\alpha}} \tag{16.14}$$

$$= \frac{1}{\lambda I_D}. \tag{16.15}$$

Thus, the intrinsic gain, $g_m r_O$, remains constant.

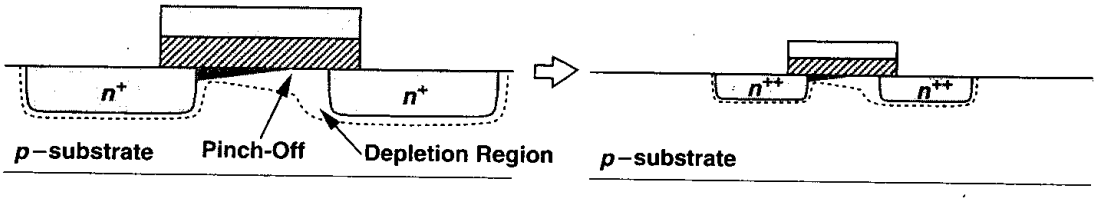


Figure 16.4 Effect of scaling on pinch-off.

The greatest impact of scaling on analog circuits is the reduction of the supply voltage. With ideal scaling, the maximum allowable voltage swings decrease by a factor of α, lowering the dynamic range[1] of the circuit. For example, if the lower end of the dynamic range is limited by thermal noise, then scaling V_{DD} by α decreases the dynamic range by

[1] Dynamic range is loosely defined as the maximum allowable voltage swing divided by the total noise voltage in the band of interest.

the same factor because g_m and hence thermal noise remain constant. Of course, since for analog circuits $(V_{DD}/\alpha)(I_{DD}/\alpha) = (V_{DD}I_{DD}/\alpha)^2$, the power dissipation drops by α^2.

In order to restore the dynamic range, the transconductance of the transistors must be increased by a factor of α^2 because thermal noise voltages and currents scale with $\sqrt{g_m}$. Thus, since voltage scaling requires that $V_{GS} - V_{TH}$ decrease by a factor of α, we note from $g_m = 2I_D/(V_{GS} - V_{TH})$ that I_D must increase by the same factor, leading to a power dissipation of $(V_{DD}/\alpha)(\alpha I_D) = V_{DD}I_D$. Also, from $g_m = \mu C_{ox}(W/L)(V_{GS} - V_{TH})$, we conclude that if C_{ox} is scaled up by α and L and $V_{GS} - V_{TH}$ are scaled down by α, then W must *increase* by α (whereas in ideal scaling it would decrease by this factor). That is, for a constant (thermal-noise limited) dynamic range, ideal scaling of linear analog circuits requires a *constant* power dissipation and a *higher* device capacitance, e.g., $(\alpha W)(L/\alpha)(\alpha C_{ox}) = \alpha WLC_{ox}$. Interestingly, if the lower end of the dynamic range is determined by kT/C noise, then to maintain a constant slew rate in switched-capacitor circuits, the bias current must scale up by a factor of α^2, resulting in an increase in the power dissipation [Problem 16.3(d)].

In practice, technology scaling has deviated from the ideal, constant-field scenario considerably. The supply voltage and MOS threshold voltage have not scaled as rapidly as device dimensions. For example, V_{DD} has decreased from 5 V to 2.5 V and V_{TH} from 0.8 V to 0.4 V as minimum channel length has dropped from 1 μm to 0.25 μm. Furthermore, many "short-channel" effects have plagued the transistors, making it difficult to obtain all of the benefits that would accrue with ideal scaling.

The reluctance of circuit designers to use a lower supply voltage and the fundamental limitations in decreasing the MOS threshold voltage have led to another scaling scenario: constant-voltage scaling. In this case, the device dimensions shrink by α, the doping levels increase by α, and the voltages remain constant, thereby increasing the electric fields by α. Such high electric fields both raise the possibility of device breakdown and exacerbate short-channel effects. In reality, technology scaling has followed a mixture of constant-field and constant-voltage trends, thus demanding innovative device design so as to achieve reliability and performance.

16.2 Short-Channel Effects

In order to appreciate the need for sophisticated device models, we briefly study some of the phenomena that manifest themselves for channel lengths below approximately 3 μm. As we will see, a basic understanding of these effects also proves essential to the design of analog (and digital) circuits.

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 in Eq. (16.5) 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.

16.2.1 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: the subthreshold behavior, variation with temperature and process, and dependence upon the channel length [6].

Let us first consider the subthreshold behavior. For long-channel devices, the subthreshold drain current can be expressed as

$$I_D = \mu C_d \frac{W}{L} V_T^2 \left(\exp \frac{V_{GS} - V_{TH}}{\zeta V_T}\right)\left(1 - \exp \frac{-V_{DS}}{V_T}\right), \tag{16.16}$$

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 $\zeta = 1 + C_d/C_{ox}$ [6]. Equation (16.16) reveals two interesting properties. First, as V_{DS} exceeds a few V_T, I_D becomes independent of the drain-source voltage and the relationship reduces to Eq. (2.30). Second, under this condition the slope of I_D on a logarithmic scale equals

$$\frac{\partial(\log_{10} I_D)}{\partial V_{GS}} = (\log_{10} e)\frac{1}{\zeta V_T}. \tag{16.17}$$

The inverse of this quantity is usually called the "subthreshold slope," S:

$$S = 2.3 V_T \left(1 + \frac{C_d}{C_{ox}}\right) \text{ V/dec}. \tag{16.18}$$

For example, if $C_d = 0.67 C_{ox}$, then $S = 100$ mV/dec, suggesting that a change of 100 mV in V_{GS} leads to a ten-fold reduction in the drain current. In order to turn off the transistor by lowering V_{GS} below V_{TH}, S must be as *small* as possible, i.e., C_d/C_{ox} 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."

The difficulty in scaling V_{TH} becomes even more serious if we take into account the variation of V_{TH} with temperature and process. The threshold voltage exhibits a temperature coefficient of approximately -1 mV/°K, yielding a 50-mV change across the commercial temperature range (0 to 50°C).[2] Process-induced variation is also in the vicinity of 50 mV, raising the margin to approximately 100 mV. Thus, it is difficult to reduce V_{TH} below several hundred millivolts.

An interesting phenomenon observed in scaled transistors is the dependence of the threshold voltage on the channel length. As shown in Fig. 16.5, 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 (Fig. 16.6). In other words, 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

[2]Interestingly, as the temperature rises, so does S, further exacerbating the situation.

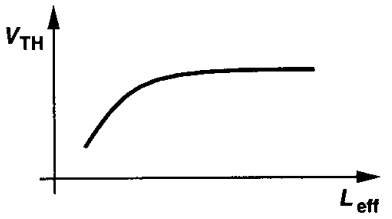


Figure 16.5 Variation of threshold with channel length.

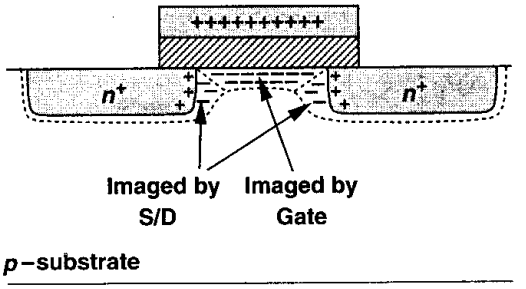


Figure 16.6 Charge sharing between source/drain depletion regions and the channel depletion region.

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}. The implication of this phenomenon in analog design is that if the length of a device is increased so as to achieve a higher output impedance, then the threshold voltage also increases by as much as 100 to 200 mV.

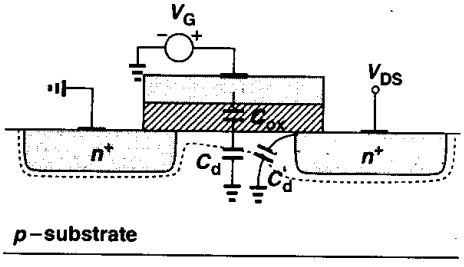
Another short-channel phenomenon related to the threshold voltage is "drain-induced barrier lowering" (DIBL). Recall from Chapter 2 that in weak inversion, as the gate voltage rises, the surface potential becomes more positive [Fig. 16.7(a)], attracting carriers from the source region. In short-channel devices, the *drain* voltage also makes the surface more positive by creating a two-dimensional field in the depletion region [6]. In essence, the drain introduces a capacitance C_d' that raises the surface potential in a manner similar to C_d. As a result, the barrier to the flow of charge and hence the threshold voltage are decreased. This effect manifests itself if the plot of Fig. 2.27 is drawn in both deep triode and saturation regions [Fig. 16.7(b)].

The principal impact of DIBL on circuit design is the degraded output impedance. This point is explained in Section 16.2.5.

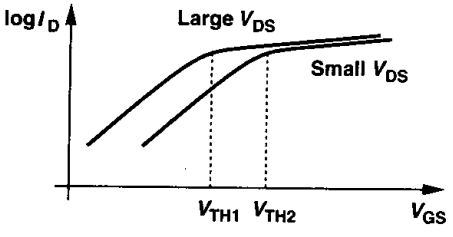
16.2.2 Mobility Degradation with Vertical Field

At large gate-source voltages, the high electric field developed between the gate and the channel confines the charge carriers to a narrower region below the oxide-silicon interface,

*While intuitive, this explanation is not quite correct. More accurate descriptions can be found in books on semiconductor devices.



(a)



(b)

Figure 16.7 (a) DIBL in a short-channel device, (b) effect of DIBL on current characteristic.

leading to more carrier scattering and hence lower mobility. Since scaling has substantially deviated from the constant-field scenario, small-geometry devices experience significant mobility degradation. An empirical equation modeling this effect is

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})}, \tag{16.19}$$

where μ_0 denotes the "low-field" mobility and θ is a fitting parameter roughly equal to $(10^{-7}/t_{ox})$ V^{-1} [7]. For example, if $t_{ox} = 100$ Å, then $\theta \approx 1$ V^{-1} and the mobility begins to fall considerably as the overdrive exceeds 100 mV. Note that θ rises as t_{ox} drops because the electric field in the oxide becomes stronger.

In addition to lowering the current capability and transconductance of MOSFETs, mobility degradation deviates the I/V characteristic from the simple square-law behavior. Specifically, whereas a square-law device generates only even harmonics in its drain current in response to a sinusoidal gate-source voltage, Eq. (16.19) predicts odd harmonics as well. In fact, writing

$$I_D = \frac{1}{2}\frac{\mu_0 C_{ox}}{1 + \theta(V_{GS} - V_{TH})}\frac{W}{L}(V_{GS} - V_{TH})^2, \tag{16.20}$$

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

$$I_D \approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[1 - \theta(V_{GS} - V_{TH})](V_{GS} - V_{TH})^2 \tag{16.21}$$

$$\approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\left[(V_{GS} - V_{TH})^2 - \theta(V_{GS} - V_{TH})^3\right]. \tag{16.22}$$

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. This is studied in Problem 16.9.

16.2.3 Velocity Saturation

The mobility of carriers also depends on the *lateral* electric field in the channel, beginning to drop as the field reaches levels of 1 V/μm. Since the carrier velocity $v = \mu E$, we note that v approaches a saturated value, about 10^7 cm/s, for sufficiently high fields. Thus, as carriers enter the channel from the source and accelerate toward the drain, they may eventually reach a saturated velocity at some point along the channel.[3] In the extreme case, where carriers experience velocity saturation along the entire channel, we can rewrite Eq. (2.2) as

$$I_D = v_{sat} Q_d \tag{16.23}$$

$$= v_{sat} W C_{ox} (V_{GS} - V_{TH}). \tag{16.24}$$

Interestingly, the current is *linearly* proportional to the overdrive voltage and does not depend on the length. In fact, as shown in Fig. 16.8, I_D-V_{DS} characteristics of devices

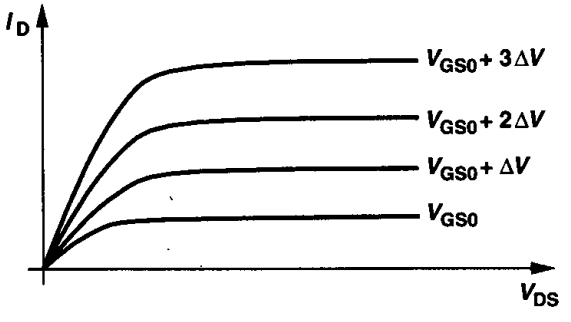


Figure 16.8 Effect of velocity saturation on drain current characteristics.

with $L < 1$ μm reveal velocity saturation because equal increments in $V_{GS} - V_{TH}$ result in roughly equal increments in I_D. We also note that $g_m = v_{sat} W C_{ox}$, concluding that the transconductance is a weak function of the drain current and channel length in the velocity-saturation regime.

Under typical bias conditions, MOSFETs experience some velocity saturation, displaying a characteristic between linear and square-law behavior. An important consequence is that, as V_{GS} increases, the drain current saturates well before pinch-off occurs. As shown in Fig. 16.9(a), carriers reach velocity saturation if V_{DS} exceeds $V_{D0} < V_{GS} - V_{TH}$, yielding a

[3] Even in long-channel devices, carriers experience velocity saturation if the drain-source voltage is high enough to pinch off the channel. At the pinch-off point, the mobile charge density is near zero, the electric field is very large, and hence the velocity of carriers is saturated.

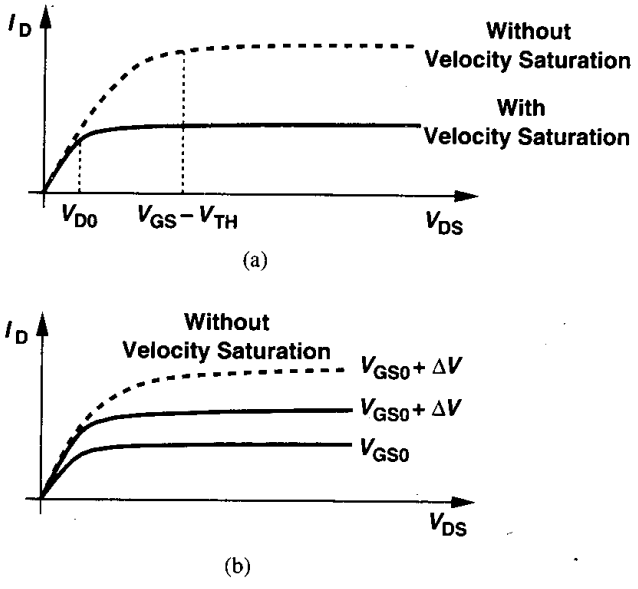


Figure 16.9 Effect of velocity saturation: (a) premature drain current saturation, (b) reduction of transconductance.

constant current quite lower than that obtained if the device saturated for $V_{DS} > V_{GS} - V_{TH}$. Furthermore, as illustrated in Fig. 16.9(b), 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.

A compact and versatile equation developed to represent velocity saturation (in the saturation region) is

$$I_D = WC_{ox}v_{sat}\frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + 2\dfrac{v_{sat}L}{\mu_{eff}}}, \tag{16.25}$$

where μ_{eff} is given by Eq. (16.19) [7, 8]. The same work provides the following equation for the drain-source voltage at the onset of premature saturation [V_{D0} in Fig. 16.9(a)]:

$$V_{DS,sat} = \frac{2\mu_{eff}L(V_{GS} - V_{TH})}{2\mu_{eff}L + V_{GS} - V_{TH}}. \tag{16.26}$$

Equation (16.25) provides two interesting results. First, if L or v_{sat} is large, the expression reduces to the square-law relationship. Second, if the *overdrive* voltage is so small that the denominator of (16.25) is approximated as $2v_{sat}L/\mu_{eff}$ and $\mu_{eff} \approx \mu_0$, then the device still follows the square-law behavior even if L is relatively small. For example, if $v_{sat} \approx 10^7$ cm/s, $L = 0.25$ μm, and $\mu_0 \approx 350$ cm^2/V/s, we have $2v_{sat}L/\mu_0 \approx 1.43$ V, recognizing that for overdrive voltages of a few hundred millivolts, the transistor operation is somewhat

close to the square law. Thus, the simplified treatment of Chapter 2 can still provide insight for many analog applications.

Equation (16.25) can be further simplified to yield additional results. Substituting for μ_{eff} from Eq. (16.19), we have

$$I_D = WC_{ox}v_{sat}\frac{(V_{GS}-V_{TH})^2}{V_{GS}-V_{TH}+\dfrac{2v_{sat}L}{\mu_0}[1+\theta(V_{GS}-V_{TH})]} \tag{16.27}$$

$$= WC_{ox}v_{sat}\frac{(V_{GS}-V_{TH})^2}{\dfrac{2v_{sat}L}{\mu_0}+\left(1+\dfrac{2v_{sat}L\theta}{\mu_0}\right)(V_{GS}-V_{TH})} \tag{16.28}$$

$$= \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\frac{(V_{GS}-V_{TH})^2}{1+\left(\dfrac{\mu_0}{2v_{sat}L}+\theta\right)(V_{GS}-V_{TH})}. \tag{16.29}$$

This equation is similar to (16.20), implying that the degradation of the mobility with both lateral and vertical fields can be represented by adding the terms $\mu_0/(2v_{sat}L)$ and θ. Thus, the results obtained from (16.20) apply here as well. For example, the drain current contains high-order nonlinear terms. Equation (16.29) can also predict the transconductance (Problem 16.10).

16.2.4 Hot Carrier Effects

Short-channel MOSFETs may experience high lateral electric fields if the drain-source voltage is large. 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 [2].

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. The substrate and gate currents are often measured to study hot carrier effects.

The scaling of technologies proceeds so as to minimize hot carrier effects. This limitation and other breakdown phenomena make the supply voltage scaling inevitable.

16.2.5 Output Impedance Variation with Drain-Source Voltage

In modeling channel-length modulation by a single constant λ, we have assumed that the output impedance of the transistor, r_O, is constant in the saturation region. In reality, however, r_O varies 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. Illustrated in Fig. 16.10, this effect is somewhat similar to the variation of the capacitance of a reversed-biased *pn* junction: with

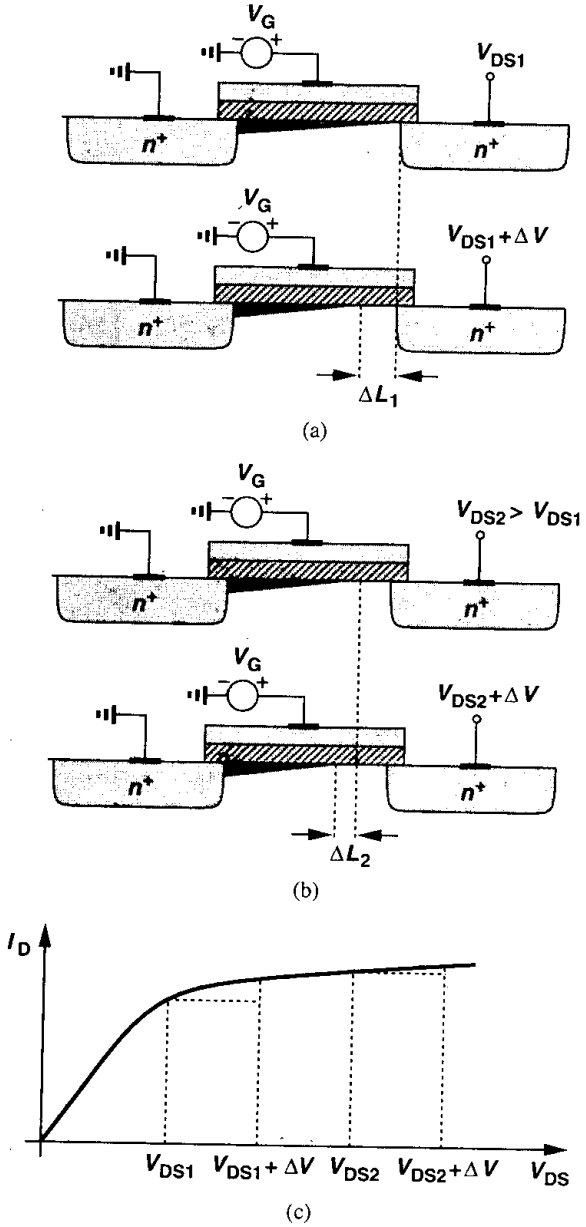


Figure 16.10 Decrement in channel length for (a) small V_{DS} and (b) large V_{DS}.

a small reverse bias, the width of the depletion region is a strong function of the voltage applied to the junction and with a large reverse bias, a weak function.

In this regime, the output impedance can be approximated as

$$r_O = \frac{2L}{1 - \dfrac{\Delta L}{L}} \frac{1}{I_D} \sqrt{\frac{qN_B}{2\epsilon_{si}}(V_{DS} - V_{DS,sat})}, \tag{16.30}$$

where $V_{D,sat}$ is the drain-source voltage at the onset of pinch-off [9]. Another approximation developed in conjunction with (16.25) and (16.26) is described in [8].

In short-channel devices, as V_{DS} increases further, drain-induced barrier lowering becomes significant, reducing the threshold voltage and increasing the drain current. This effect roughly cancels that expressed by (16.30), 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 the output impedance. The overall behavior of r_O is plotted in Fig. 16.11.

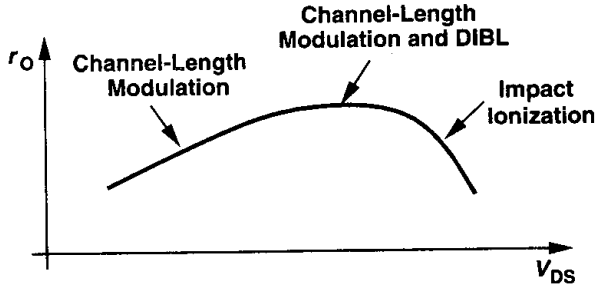


Figure 16.11 Overall variation of output resistance as a function of V_{DS}.

The variation of r_O gives rise to nonlinearity in many circuits. In a cascode op amp, for example, as the output voltage varies, so does the output impedance of the cascode devices and hence the voltage gain of the circuit. Furthermore, impact ionization limits the maximum gain that can be obtained from cascode structures because it introduces a small-signal resistance from the drain to the *substrate* rather than to the source.

16.3 MOS Device Models

Since the introduction of the first MOS model in the mid-1960s [10], tremendous research has been expended on improving the accuracy of models as device dimensions scale down. Developed between the mid-1960s and the late 1970s, the Level 1, 2, and 3 models consecutively included higher order effects so as to provide reasonable accuracy with respect to measured transistor characteristics for channel lengths as small as 1 μm. Following this set were the Compact Short-Channel IGFET Model (CSIM) from AT&T Bell Laboratories and the Berkeley Short-Channel IGFET Model (BSIM) from University of California, Berkeley

in the mid-1980s. These models proved inadequate for analog design and were followed by BSIM2, HSPICE level 28, BSIM3, and a number of others in the late 1980s and early 1990s.

MOS device modeling continues to pose a challenge—especially for high-frequency operation—because even today's sophisticated models become inadequate after one or two technology generations (e.g., from 0.5 μm to 0.35 μm to 0.25 μm). Our objective is to develop a basic understanding of some of the models to the extent necessary for simulations. We should also mention that 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. The interested reader is referred to [1] for an in-depth coverage.

16.3.1 Level 1 Model

Also known as the Shichman and Hodges Model [10], this representation uses the parameters listed in Table 2.1 and is based on the following equations:

$$I_D = \frac{1}{2}K_P\frac{W}{L-2L_D}[2(V_{GS}-V_{TH})V_{DS}-V_{DS}^2](1+\lambda V_{DS}) \quad \text{Triode Region} \tag{16.31}$$

$$I_D = \frac{1}{2}K_P\frac{W}{L-2L_D}(V_{GS}-V_{TH})^2(1+\lambda V_{DS}) \quad \text{Saturation Region} \tag{16.32}$$

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.

The device capacitances are represented according to the simple model described in Chapter 2, but with one modification. Since in that model, C_{GS} abruptly changes from $(2/3)WLC_{ox} + WC_{ov}$ in saturation to $(1/2)WLC_{ox} + WC_{ov}$ in the triode region [and C_{GD} from WC_{ov} to $(1/2)WLC_{ox} + WC_{ov}$], most computation algorithms experience convergence difficulties here. For this reason, C_{GS} and C_{GD} in the triode region are formulated as

$$C_{GS} = \frac{2}{3}WLC_{ox}\left\{1-\frac{(V_{GS}-V_{DS}-V_{TH})^2}{[2(V_{GS}-V_{TH})-V_{DS}]^2}\right\} + WC_{ov} \tag{16.33}$$

$$C_{GD} = \frac{2}{3}WLC_{ox}\left\{1-\frac{(V_{GS}-V_{TH})^2}{[2(V_{GS}-V_{TH})-V_{DS}]^2}\right\} + WC_{ov} \tag{16.34}$$

$$C_{GB} = 0. \tag{16.35}$$

We note that if the device operates at the edge of saturation, $V_{GS} - V_{DS} = V_{TH}$, $C_{GS} = (2/3)WLC_{ox} + WC_{ov}$, and $C_{GD} = WC_{ov}$. Thus, the capacitance values change continuously from one region to another.

The Level 1 model maintains reasonable I/V accuracy for channel lengths as small as roughly 4 μm, but it still predicts the output impedance of transistors in saturation quite poorly.

16.3.2 Level 2 Model

The Level 1 model began to manifest its shortcomings as channel lengths fell below approximately 4 μm. The Level 2 model was then developed to represent many high-order effects.

An assumption that we made in Chapter 2 in deriving the square-law characteristics was a constant threshold voltage along the channel. This assumption is not correct even for long-channel devices because the charge in the depletion region under the channel varies according to the local voltage (Fig. 16.12). Since the inversion layer and the depletion region

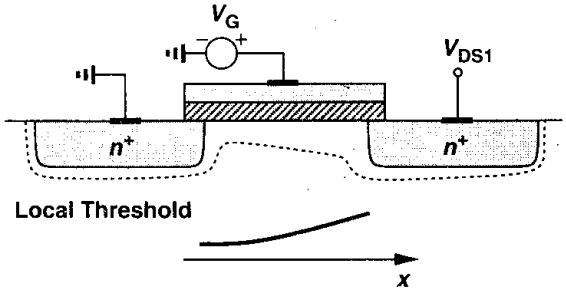


Figure 16.12 Variation of threshold along the channel.

must image the charge on the gate, as the inversion layer vanishes in the direction toward the drain, the depletion region must enclose more charge. Performing the integration in Section 2.2.2 with a varying threshold voltage yields [1]:

$$I_D = \mu C_{ox} \frac{W}{L} \left\{ (V_{GS} - V_{TH0}) V_{DS} - \frac{V_{DS}^2}{2} - \frac{2}{3}\gamma \left[(V_{DS} - V_{BS} + 2\phi_F)^{3/2} - (-V_{BS} + 2\phi_F)^{3/2} \right] \right\}. \quad (16.36)$$

Interestingly, even for $V_{BS} = 0$, I_D exhibits some dependence on γ. 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 [1]:

$$V_{D,sat} = V_{GS} - V_{TH0} - \phi_F + \gamma^2 \left[1 - \sqrt{1 + \frac{2}{\gamma^2}(V_{GS} - V_{TH0} + \phi_F)} \right]. \quad (16.37)$$

In the saturation region, the drain current is

$$I_{DS} = I_{D,sat} \frac{1}{1 - \lambda V_{DS}}, \quad (16.38)$$

where $I_{D,sat}$ is calculated from (16.36) for $V_{DS} = V_{DS,sat}$.

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. In the Level 2 implementation, if λ is not specified, it is obtained by calculating the width of the depletion region between the pinch-off point and the edge of the drain. Using simple relationships for the depletion region of a *pn* junction, we can write

$$\Delta L = \sqrt{\frac{2\epsilon_{si}}{qN_{sub}}[\phi_B + (V_{DS} - V_{D,sat})]}, \tag{16.39}$$

where $V_{D,sat}$ denotes the pinch-off voltage.[4]

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 [1]! To resolve this issue, ΔL is actually obtained by a "fixed-up" equation:

$$\Delta L = \sqrt{\frac{2\epsilon_{si}}{qN_{sub}}\left(V_1 + \sqrt{1 + V_1^2}\right)}, \tag{16.40}$$

where $V_1 = (V_{DS} - V_{D,sat})/4$. The channel-length modulation coefficient is then expressed as $\lambda = \Delta L/(LV_{DS})$. An attribute of (16.40) is that the output conductance of the transistor varies as V_{DS} increases, an effect not represented by the first-order model using a constant λ.

The Level 2 model also includes the degradation of the mobility with the vertical field in the channel. The mobility is calculated from

$$\mu_s = \mu_0 \left(\frac{\epsilon_{si}}{C_{ox}} \cdot \frac{U_c}{V_{GS} - V_{TH} - U_t V_{DS}}\right)^{U_e}, \tag{16.41}$$

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.

The subthreshold behavior implemented in the Level 2 model defines a voltage V_{on} as $V_{on} = V_{TH} + \zeta V_T$, 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

$$I_{DS} = I_{on} \exp \frac{V_{GS} - V_{on}}{\zeta V_T}, \tag{16.42}$$

where I_{on} is the drain current calculated in strong inversion [Eq. (16.36)] for $V_{GS} = V_{on}$. An important drawback of this representation is the discontinuity in the slope of I_D as the device goes from the subthreshold region to strong inversion (Fig. 16.13), leading to various difficulties and errors in simulation.

In addition to the above effects, the Level 2 model represents two other short-channel phenomena: the variation of V_{TH} with L, and velocity saturation. The implementation of these effects is quite involved and can be found in [1].

Measured data [1] indicate that the Level 2 model provides reasonable I/V accuracy for wide, short devices in the saturation region with $L \approx 0.7\ \mu$m but it suffers from substantial

[4]The junction is considered "one-sided" here, i.e., the drain doping level is much higher.

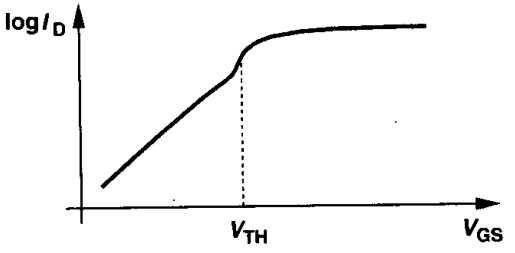


Figure 16.13 Kink in drain current characteristic in Level 2 model.

error in representing the output impedance and the transition point between saturation and triode regions. For narrow or long devices, the model is quite inaccurate.

16.3.3 Level 3 Model

The Level 3 model realization is somewhat similar to the Level 2 model, with some equations simplified and many empirical constants introduced to improve the accuracy for channel lengths as small as 1 μm.

This model expresses the threshold voltage as

$$V_{TH} = V_{TH0} + F_s\gamma\sqrt{2\phi_F - V_{BS}} + F_n(2\phi_F - V_{BS}) + \xi\frac{8.15 \times 10^{-22}}{C_{ox}L_{eff}^3}V_{DS}, \quad (16.43)$$

where F_s and F_n represent short-channel and narrow-channel effects,[5] respectively, and ξ models drain-induced barrier lowering.

The mobility equation involves both vertical and lateral field effects and is expressed as:

$$\mu_1 = \frac{\mu_{eff}}{1 + \dfrac{\mu_{eff}V_{DS}}{v_{max}L_1}}, \quad (16.44)$$

where

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})}, \quad (16.45)$$

and v_{max} denotes the maximum velocity of the carriers in the channel. As can be seen from (16.44) and (16.45), μ_{eff} models the effect of the vertical field while μ_1 adds that of the lateral field as well.

The drain current is realized as:

$$I_D = \mu_1 C_{ox}\frac{W_{eff}}{L_{eff}}\left[V_{GS} - V_{TH0} - \left(1 + \frac{F_s\gamma}{4\sqrt{2\phi_F - V_{BS}}} + F_n\right)\frac{V'_{DS}}{2}\right]V'_{DS}, \quad (16.46)$$

where $V'_{DS} = V_{D,sat}$ if the device is in saturation. The quantity $V_{D,sat}$ represents both channel pinch-off and velocity saturation (Fig. 16.9) and is expressed by relatively complex

[5] For narrow-channel devices, the threshold voltage *increases* if the *width* is reduced [6].

equations [1]. The subthreshold current relations are similar to those of the Level 2 model, still suffering from derivative discontinuity near strong inversion.

The Level 3 model employs more sophisticated methods of computing channel-length modulation as well as charge and capacitance parameters. The details can be found in [1]. Comparison with measured data [1] suggests that the 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. Shown in Fig. 16.14 for a short-channel device, the variation of r_O with V_{DS} is quite poorly modeled.

Figure 16.14 Kink in output resistance in Level 3 model.

16.3.4 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. However, as transistors were scaled to submicron dimensions, it became increasingly more difficult to introduce physically meaningful equations that would be both accurate and computationally efficient. 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. The general expression is of the form:

$$P = P_0 + \frac{\alpha_P}{L_{eff}} + \frac{\beta_P}{W_{eff}}, \tag{16.47}$$

where P_0 is the value of the parameter for a long, wide transistor ($P = P_0$ if $L_{eff}, W_{eff} \to \infty$), and α_P and β_P are fitting factors. For example, the mobility is computed as:

$$\mu = \mu_0 + \frac{\alpha_\mu}{L_{eff}} + \frac{\beta_\mu}{W_{eff}}. \tag{16.48}$$

The formulation of (16.47) nonetheless becomes less accurate at small dimensions [1].

The device equations and fitting parameters used in BSIM are beyond the scope of this book. Using approximately 50 parameters, this model provides the following improvements over the Level 3 version [1]: (1) the dependence of mobility upon the vertical field includes

the substrate voltage; (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.

Measured results in a 0.7-μm technology [1] indicate that BSIM avoids gross errors in the I/V characteristics for various device dimensions, but its accuracy for narrow, short transistors is somewhat poor.

In addition to shortcomings at channel lengths below approximately 0.8 μm, BSIM suffers from other subtle inaccuracies. For example, at large drain-source voltages, BSIM predicts a *negative* output resistance for saturated MOSFETs. Furthermore, in deep triode region, BSIM still exhibits slight discontinuities in the drain current [1].

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. Nevertheless, measured results indicate that the overall accuracy of the model is only marginally higher than that of BSIM. For short, narrow transistors, BSIM2 suffers from large errors in the triode region and even substantial "kinks" in the saturation region [1].

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. Parameter extraction, modeling process variations, and the need for extensive use of polynomials made the generation and application of these models quite difficult. 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. For channel lengths as low as 0.25 μm, BSIM3 provides reasonable accuracy for subthreshold and strong inversion operation while still suffering from large errors in predicting the output impedance.

16.3.5 Other Models

In addition to the Level 1–3 models and the three generations of BSIM, a number of other MOS models have been introduced. Among these, HSPICE Level 28, MOS9, and the Enz-Krummenacher-Vittoz (EKV) model are the most notable, for they provide new approaches to representing the behavior of MOSFETs [1]. For example, the HSPICE Level 28 model improves the dependence of accuracy upon device dimensions by expressing the parameters as:

$$P = P_0 + \alpha\left(\frac{1}{L} - \frac{1}{L_{ref}}\right) + \beta\left(\frac{1}{W} - \frac{1}{W_{ref}}\right) + \gamma\left(\frac{1}{L} - \frac{1}{L_{ref}}\right)\left(\frac{1}{W} - \frac{1}{W_{ref}}\right), \tag{16.49}$$

where L_{ref} and W_{ref} denote the dimensions of a "reference" device, i.e., a transistor whose characteristics have been measured. Thus, the dependence is expressed in terms of *increments* with respect to characterized transistors rather than the absolute value of the

dimensions, yielding a potentially higher accuracy. Also, the term proportional to the product of the length and width increments facilitates curve fitting.

The EKV model [11] substantially departs from traditional views of MOSFET operation by considering the *bulk*, rather than the source, as the reference point for all voltages. This approach thus avoids distinguishing between the source and drain terminals and, more importantly, introduces a single drain-source current equation that is valid for both subthreshold and saturation regions.

The reader is referred to [1] for an extensive study of these models.

16.3.6 Charge and Capacitance Modeling

The simple gate capacitance model described in Chapter 2 for the Level 1 model, called the Meyer capacitance model [1], 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. For example, as illustrated in Fig. 16.15, a periodic

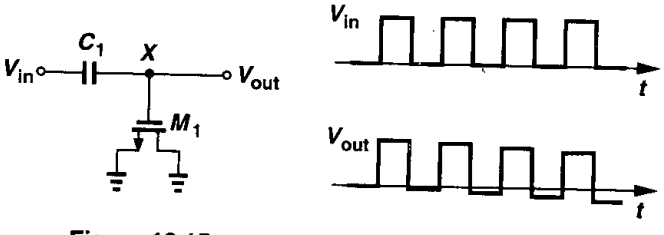


Figure 16.15 Annihilation of charge in simulation.

rectangular waveform applied to a voltage 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. This effect arises from the calculation of charge by integrating capacitor voltages with respect to time, an operation that accumulates small errors in the simulation.[6] 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 relates to partitioning of the channel charge between the source and drain terminals. The assumption that in the triode region, $C_{GS} = C_{GD} = (1/2)WLC_{ox} + WC_{ov}$, and in the saturation region, $C_{GS} = (2/3)WLC_{ox} + WC_{ov}$ and $C_{GD} = WC_{ov}$ is quite inaccurate for short-channel devices, requiring flexible partitioning for ease of curve fitting. In BSIM and BSIM3, for example, three different charge partitioning scenarios (40%/60%, 50%/50%, and 0%/100%) are available.

Recent efforts have created more sophisticated charge and capacitance models for MOS devices so as to improve the accuracy, especially for analog applications. However, as with many other modeling improvements, the resulting equations are quite cumbersome, imparting little intuition. The reader is referred to [1] for details.

---

[6] Another source of error here is the assumption that the device capacitances are reciprocal, e.g., $C_{GS} = C_{SG}$ [1].

16.3.7 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.,

$$E_g = 1.16 - \frac{7.02 \times 10^{-4} T^2}{T + 1108}, \tag{16.50}$$

and

$$\mu = \mu_0 \left(\frac{300}{T}\right)^{3/2}, \tag{16.51}$$

where $\mu_0 = \mu(T = 300° \text{ K})$.

BSIM3 incorporates a few more parameters to represent the temperature dependence of phenomena such as velocity saturation and the effect of subthreshold voltage on V_{TH}. It is unclear at this point how accurately BSIM3 expresses the temperature variation of MOS devices and circuits.

16.4 Process Corners

Unlike bipolar transistors, MOSFETs suffer from substantial parameter variations from wafer to wafer and from lot to lot. Despite decades of technology advancement, the large variability of CMOS circuits remains a fact with which digital and analog designers must cope.

In order to facilitate the task of circuit design to some extent, process engineers guarantee a performance envelope for the devices, in essence tightening the anticipated parameter variations by discarding wafers that fall out of the envelope (Fig. 16.16). Of course, in

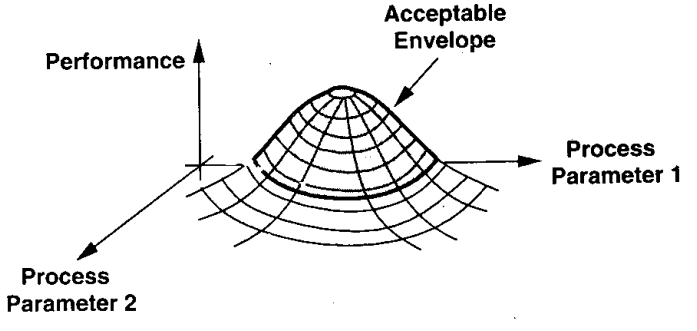


Figure 16.16 Performance envelope as a function of process parameters.

their eternal battle, circuit designers insist on a tighter variability space so that they can design more aggressively whereas process engineers tend to enlarge the envelope as much as possible so as to increase the yield. For example, it is common in today's CMOS technologies to obtain a gate delay that varies by a factor of two to one with process and temperature.

The performance envelope furnished to designers has traditionally been one suited to digital circuits and constructed in the form of "process corners." Illustrated in Fig. 16.17,

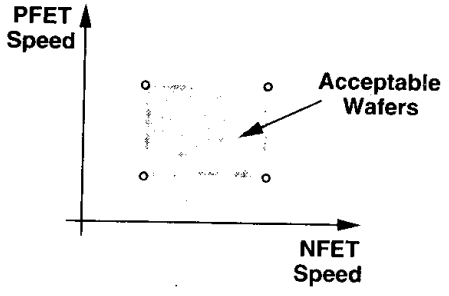


Figure 16.17 Process corners based on speed of NMOS and PMOS devices.

the idea is to constrain the speed envelope of the NMOS and PMOS transistors to a rectangle defined by four corners: fast NFET and fast PFET; slow NFET and slow PFET; fast NFET and slow PFET; and slow NFET and fast PFET. For example, transistors having a thinner gate oxide and lower threshold voltage fall near the fast corner. The device models corresponding to each corner are extracted from wafers whose NMOS or PMOS test structures display a large or small gate delay, and the actual corners are chosen so as to obtain an acceptable yield. Thus, only wafers satisfying these specifications are considered acceptable. Simulation of circuits for various process corners and temperature extremes is essential to determining the yield.

16.5 Analog Design in a Digital World

Memories and processors constitute the major portion of today's semiconductor business. Thus, as explained in Chapters 17 and 18, most CMOS technologies are designed, optimized, and *characterized* for digital applications. Despite the increasing emphasis on the "analog" accuracy of device models, we are still far from a point where we can fully trust the absolute numbers obtained in circuit simulations. Analog designers routinely encounter discrepancies in SPICE, for example, between ac analysis and transient analysis. Moreover, many device models fail simple benchmark tests [12] and effects such as flicker (and thermal) noise and mismatch require measured data before they can be accurately reflected in simulations. Subtle, yet important phenomena such as nonlinearity of the device output resistance are represented incorrectly even in the most recent models. Also, the device models extracted from a wafer often fail to accurately predict the speed of the circuits fabricated on the same wafer! These difficulties are intensified by the rapid migration of CMOS technologies from one generation to the next.

Under these conditions, analog design relies on experience, intuition, and *measured* data. In fact, the design of complex, high-performance analog circuits may require data points that can be obtained only by first fabricating and characterizing many simpler test circuits [13].

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume $V_{DD} = 3$ V where necessary. Also, assume all transistors are in saturation.

16.1. Silicon dioxide breaks down at high electric fields. Explain what happens if ideal scaling is performed while keeping the gate oxide thickness constant.

16.2. The maximum doping level that can be established in the source and drain regions is limited by the "solid solubility" of silicon. Explain what happens to the S/D junction capacitance and series resistance as ideal scaling occurs but the S/D doping level remains constant. Does DIBL become more or less significant?

16.3. Suppose the supply voltage of a switched-capacitor amplifier is reduced by a factor of two and so is the maximum allowable output voltage swing. In order to maintain the dynamic range constant, the noise voltage must scale down by the same factor.

(a) If the noise is only of kT/C type, how should the capacitors in the circuit be scaled?

(b) If the time constant is given by G_m/C, where G_m denotes the transconductance of a one-stage op amp, how should G_m be scaled to maintain the same small-signal time constant?

(c) How should the dimensions and tail current of the input differential pair of the op amp be scaled?

(d) Repeat parts (b) and (c) where the slew rate must remain constant.

16.4. Explain how each parameter in Eq. (16.16) scales in an ideal constant-field scaling scenario. What happens to the subthreshold slope?

16.5. A common-gate stage designed for an input impedance of 50 Ω undergoes ideal scaling. If $\lambda = \gamma = 0$, what is the input impedance?

16.6. Repeat Problem 16.5 if $\lambda \neq 0$, $\gamma \neq 0$, and the load is a MOS current source that is also scaled.

16.7. For power-conscious applications, a figure of merit is defined as the transconductance of devices normalized to their bias current. Determine this quantity for long-channel devices operating in strong inversion or the subthreshold region. At what drain current are these two equal?

16.8. Explain why the mobile charge density cannot drop to exactly zero at any point along the channel. What happens beyond the pinch-off point?

16.9. Using Eq. (16.20), calculate the transconductance of a MOSFET. What happens if the overdrive voltage is very small or very large?

16.10. Using Eq. (16.29), calculate the transconductance of a MOSFET. Prove that

$$g_m = \frac{I_D}{V_{GS} - V_{TH}} \left[1 + \frac{1}{1 + \left(\frac{\mu_0}{2 v_{sat} L} + \theta \right)(V_{GS} - V_{TH})} \right]. \tag{16.52}$$

16.11. Suppose the channel-length modulation coefficient λ is modified as $\lambda/(1 + \kappa V_{DS})$, where κ is a constant, to represent the dependence of the output impedance upon V_{DS}. Calculate r_O. Explain how a current source with such behavior introduces distortion in the voltage across it.

16.12. Assuming the devices in Fig. 16.18 experience complete velocity saturation, derive expressions for the voltage gain of each circuit in terms of W and v_{sat}. Assume $\lambda = \gamma = 0$.

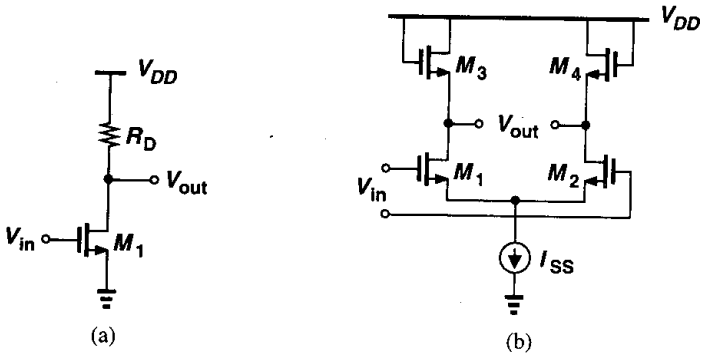


Figure 16.18

16.13. Using Eq. (16.36), calculate g_{mb} and compare the result with that derived in Chapter 2.

16.14. From Eq. (16.50), determine $\partial E_g / \partial T$ at room temperature and explain how it affects bandgap reference voltages.

16.15. Suppose the fast corners of a process result from a higher μC_{ox}. Explain what happens to the voltage gain and the input thermal noise of the circuits shown in Fig. 16.19 at the four corners of the process if the transistors are biased at a constant current in saturation.

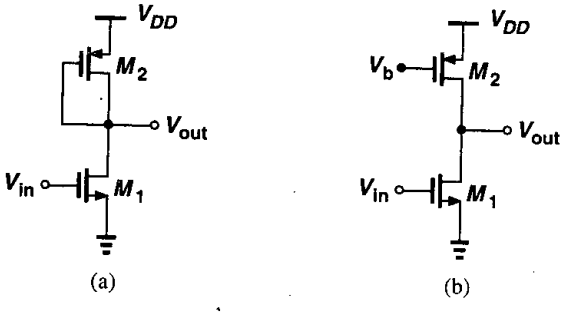


Figure 16.19

16.16. Repeat Problem 16.15 if each transistor is biased with a fixed V_{GS}.

References

1. D. P. Foty, *MOSFET Modeling with SPICE,* Upper Saddle River, NJ: Prentice-Hall, 1997.
2. Y. Tsividis, *Operation and Modeling of the MOS Transistor,* Second Ed., New York: McGraw-Hill, 1999.
3. P. Antognetti and G. Massobrio, Editors, *Semiconductor Device Modeling with SPICE,* New York, McGraw-Hill, 1988.

4. R. H. Dennard et al., "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. of Solid-State Circuits,* vol. 9, pp. 256–268, Oct. 1974.
5. G. E. Moore, "Progress in Digital Integrated Circuits," *IEDM Tech. Dig.,* pp. 11–14, Dec. 1975.
6. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices,* New York: Cambridge University Press, 1998.
7. C. G. Sodini, P. K. Ko, and J. L. Moll, "The Effect of High Fields on MOS Device and Circuit Performance," *IEEE Tran. on Electron Devices,* vol. 31, pp. 1386–1393, Oct. 1984.
8. P. K. Ko, "Approaches to Scaling," pp. 1–35, in *Advanced MOS Device Physics,* N.G. Einspruch and G. Gildenblat, Editors, San Diego: Academic Press, 1998.
9. S. Wong and A. T. Salama, "Impact of Scaling on MOS Analog Performance," *IEEE J. of Solid-State Circuits,* vol. 18, pp. 106–114, Feb. 1983.
10. H. Shichman and D. A. Hodges, "Modeling and Simulation of Insulated Field Effect Transistor Switching Circuits, *IEEE J. of Solid-State Circuits,* vol. 3, pp. 285–289, Sept. 1968.
11. C. C. Enz, F. Krummenacher, and E. Vittoz, "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low Voltage and Low Current Applications," *Analog Integrated Circuits and Signal Processing,* vol. 8, pp. 83–114, 1995.
12. Y. Tsividis and K. Suyama, "MOSFET Modeling for Analog Circuit CAD: Problems and Prospects," *IEEE J. of Solid-State Circuits,* vol. 29, pp. 210–216, March 1994.
13. B. Razavi, "CMOS Technology Characterization for Analog and RF Design," *IEEE J. of Solid-State Circuits,* vol. 34, pp. 268–276, March 1999.

Chapter 17

CMOS Processing Technology

With the high-order effects of MOS devices covered in Chapter 16, we now study the fabrication of CMOS technologies. A solid understanding of device processing proves essential in the design and layout of ICs because many limitations imposed on the performance of circuits are related to fabrication issues. Furthermore, today's semiconductor technology demands that process engineers and circuit designers interact regularly so as to understand each other's needs, necessitating a good knowledge of each discipline.

In this chapter, we deal with the processing technology of CMOS devices, aiming to provide a simple view of the fabrication steps and their relevance to circuit design and layout. We begin with a brief description of basic fabrication steps such as wafer processing, photolithography, oxidation, ion implantation, deposition, and etching. Next, we study the fabrication sequence of MOS transistors in detail. Finally, we describe the processing of passive devices and interconnections.

17.1 General Considerations

Before delving into a detailed study of fabrication, it is instructive to consider the basic structure of NMOS and PMOS transistors and predict the required processing steps. As shown in Fig. 17.1, a p-type substrate (wafer) serves as the foundation upon which n-wells, source/drain regions, gate dielectric, polysilicon, n-well and substrate ties, and metal interconnects are built. Considering both the side view and the top view, we may raise the following questions: (1) How are various regions defined so accurately? For example, how is a gate polysilicon line with a minimum dimension of 0.25 μm fabricated while maintaining a distance of 0.25 μm from another polysilicon line? (2) How are the n-wells and S/D regions built? (3) How are the gate oxide and polysilicon fabricated? (4) How are the gate oxide and polysilicon *aligned* with the S/D regions? (5) How are the contact windows created? (6) How are the metal interconnect layers deposited?

Modern CMOS technologies involve more than 200 processing steps, but for our purposes, we can view the sequence as a combination of the following operations: (1) wafer processing to produce the proper type of substrate; (2) photolithography to precisely define

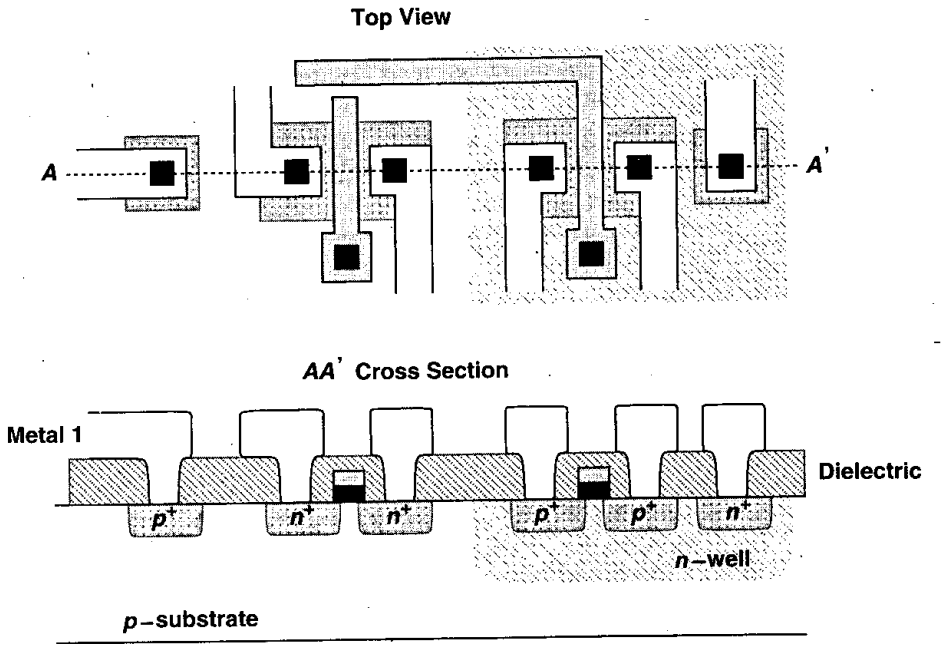


Figure 17.1 Side view and top view of MOS devices.

each region; (3) oxidation, deposition, and ion implantation to *add* materials to the wafer; (4) etching to *remove* materials from the wafer. Many of these steps require "heat treatment," i.e., the wafer must undergo a thermal cycle inside a furnace.

In semiconductor processing and characterization, we often refer to the "sheet resistance" of a layer. The total resistance of a rectangular bar is $R = \rho L/(W \cdot t)$, where ρ is the resistivity of the material, and L, W, and t, denote the length, width, and thickness of the bar, respectively. In integrated circuits, the resistivity and thickness of the layers are set by fabrication materials and processing steps and cannot be changed in the layout. The quantity $R_\square = \rho/t$ is thus defined as the sheet resistance, combining two constants of the technology. Since $R = R_\square$ for $W = L$, i.e., for a square geometry, we express $R_\square$ in terms of ohms per square. For example, for a sheet resistance of 10 $\Omega/\square$, a geometry with $W = 2$ μm and $L = 20$ μm has a resistance of $R = 10\ \Omega/\square \times (20/2) = 100\ \Omega$. In fact, we may say "this line is 10 squares long," meaning that $L/W = 10$ and $R = 10R_\square$.

17.2 Wafer Processing

The starting wafer in a CMOS technology must be created with a very high quality. That is, the wafer must be grown as a single-crystal silicon body having a very small number of "defects," e.g., dislocations in the crystal or unwanted impurities. Furthermore, the wafer must contain the proper type and level of doping so as to achieve the required resistivity.

This is accomplished by the "Czochralski method," whereby a seed of crystalline silicon is immersed in molten silicon and gradually pulled out while rotating. As a result, a large single-crystal cylindrical "ingot" is formed that can be sliced thin into wafers. The diameter of the wafer has scaled up with new technology generations, exceeding 20 cm (8 in) today. Note that dopants are added to the molten silicon to obtain the desired resistivity. The wafers are then polished and chemically etched, thereby removing damages on the surface that are created during slicing. In most CMOS technologies, the wafer has a resistivity of 0.05 to 0.1 Ω-cm and a thickness of approximately 500 to 1000 μm (which is reduced to a few hundred microns after all of the processing steps).

17.3 Photolithography

Photolithography, or simply lithography, is the first step in transferring the circuit layout information to the wafer. As shown in the top view of Fig. 17.1 and explained in Chapter 18 in more detail, the layout consists of polygons representing different types of "layers," e.g., n-well, S/D regions, polysilicon, contact windows, etc. For fabrication purposes, we decompose the layout into these layers. For example, the layout of Fig. 17.1 can be viewed as five different layers shown in Fig. 17.2, each of which must be created on the wafer with a very high precision. Note that the "active" (or "diffusion") layer includes the source/drain regions and the p^+ and n^+ openings serving as the substrate and well ties.

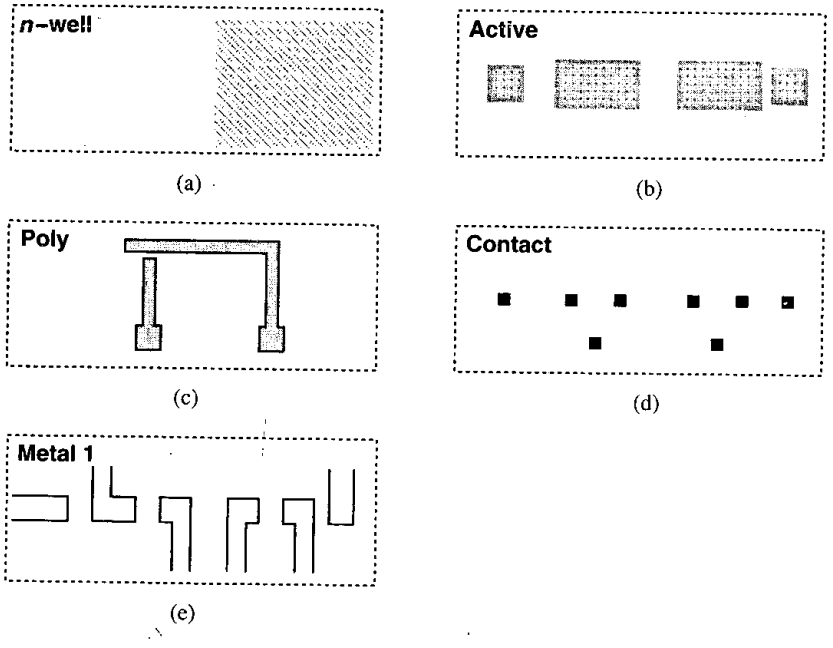


Figure 17.2 Layers comprising the structures of Fig. 17.1.

To understand how a layer is transferred from the layout to the wafer, let us consider the n-well pattern of Fig. 17.2(a) as an example. This pattern is "written" to a transparent glass "mask" by a precisely controlled electron beam [Fig. 17.3(a)]. Also, as depicted in Fig. 17.3(b), the wafer is covered by a thin layer of "photoresist," a material whose etching properties change upon exposure to light.[1] Subsequently, the mask is placed on top of the wafer and the pattern is projected onto the wafer by ultraviolet (UV) light [Fig. 17.3(c)]. The photoresist "hardens" in the regions exposed to light and remains "soft" under the opaque rectangle. The wafer is then placed in an etchant that dissolves the "soft" photoresist area, thereby exposing the silicon surface [Fig. 17.3(d)]. Now, an n-well can be created in the exposed area. We call this set of operations a lithography sequence.

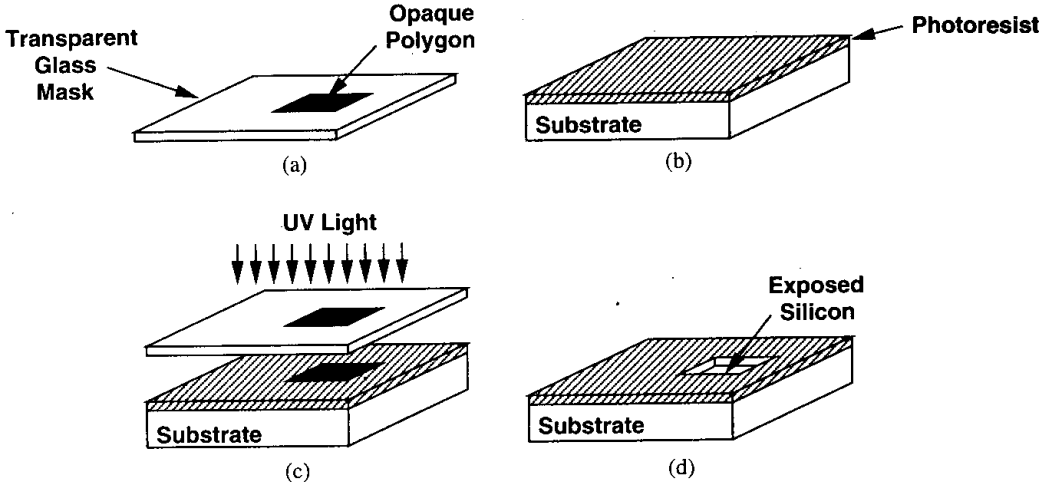


Figure 17.3 (a) Glass mask used in lithography, (b) coverage of wafer by photoresist, (c) selective exposure of photoresist to UV light, (c) exposed silicon after etching.

In summary, the sequence associated with the lithography of each layer involves one mask and three processing steps: (1) cover wafer with photoresist; (2) align mask on top and expose to light; (3) etch exposed photoresist. The example of Fig. 17.2 therefore requires at least five masks and hence five lithography sequences.

We should mention that two types of photoresists are used in processing. A "negative" photoresist hardens in the areas exposed to light and a "positive" photoresist hardens in the areas not exposed to light. As explained later in this chapter, both types prove useful in fabrication.

The number of masks in a process heavily impacts the overall cost of fabrication, eventually influencing the unit price of the chip. This is so for two reasons: each mask costs several thousand dollars, and, owing to the necessary precision, lithography is a slow and expensive task. In fact, CMOS technology originally became attractive by virtue of the relatively small number of masks—about seven—that it required. Although in modern CMOS processes,

[1] In practice, a thin layer of oxide is grown before depositing the photoresist to protect the surface.

this number is close to 25 (and the total cost of masks greater than $200,000), the cost of each IC has nonetheless remained low because both the number of transistors per unit area and the size of the wafer have steadily increased.

17.4 Oxidation

A unique property of silicon is that it can produce a very uniform oxide layer on the surface with little strain in the lattice, allowing the fabrication of gate oxide layers as thin as a few tens of angstroms (only several *atomic* layers). In addition to serving as the gate dielectric, silicon dioxide can act as a protective coating in many steps of fabrication. Also, in areas between the devices, a thick layer of SiO_2, called the "field oxide" (FOX) is grown, providing the foundation for interconnect lines that are formed in subsequent steps (Fig. 17.4).

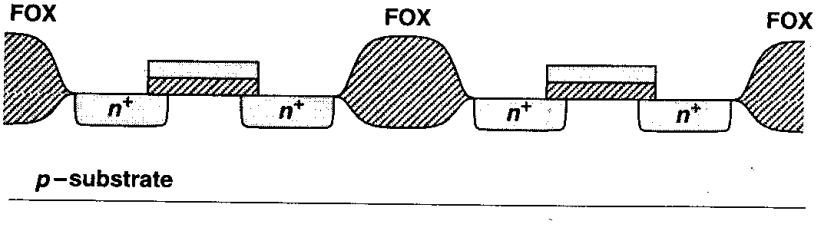


Figure 17.4 Field oxide.

Silicon dioxide is "grown" by placing the exposed silicon in an oxidizing atmosphere such as oxygen at a temperature around 1000°C. The rate of growth depends on the type and pressure of the atmosphere, the temperature, and the doping level of the silicon.

The growth of the gate oxide is a very critical step in the process. Since the oxide thickness, t_{ox}, determines both the current handling and reliability of the transistors, it must be controlled to within a few percent. For example, the oxide thicknesses of two transistors separated by 20 cm on a wafer must differ by less than a few angstroms, requiring extremely high uniformity across the wafer and hence a slow growth of the oxide. Also, the "cleanness" of the silicon surface under the oxide affects the mobility of the charge carriers and thus the current drive, transconductance, and noise of the transistors.

17.5 Ion Implantation

In many steps of fabrication, dopants must be selectively introduced into the wafer. For example, after the lithography sequence of Fig. 17.3 is completed, the n-well is formed by entering dopants into the exposed silicon area. Similarly, the source and drain regions of transistors require selective addition of dopants to the wafer.

The most common method of introducing dopants is "ion implantation," whereby the doping atoms are accelerated as a high-energy focused beam, hitting the surface of the wafer and penetrating the exposed areas [Fig. 17.5(a)]. The doping level (dosage) is determined by the intensity and duration of the implantation, and the depth of the doped region is set

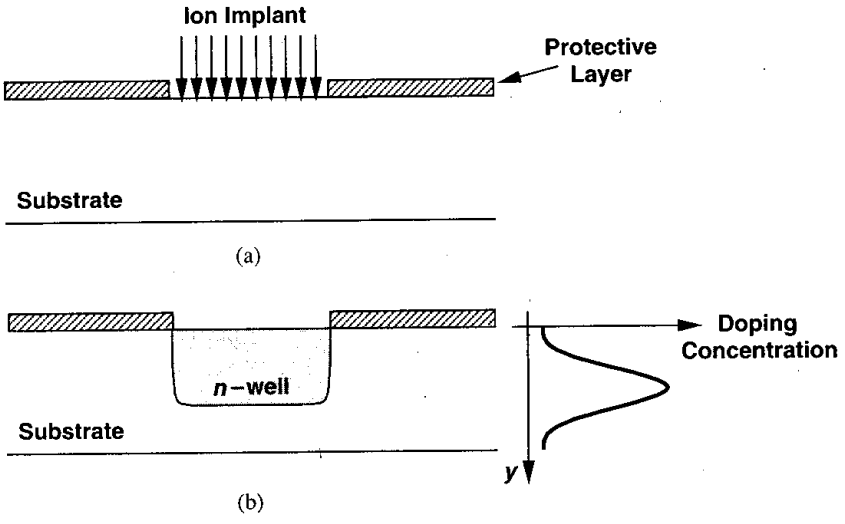


Figure 17.5 (a) Ion implantation, (b) retrograde profile.

by the energy of the beam. As shown in Fig. 17.5, with a high energy, the peak of the doping concentration in fact occurs well below the surface, thereby creating a "retrograde" profile. Such a profile is desirable for the n-well because it establishes a low resistivity near the bottom, reducing susceptibility to latch-up (Section 17.8), and a low doping level at the surface, decreasing the S/D junction capacitance of PMOS devices.

Another important application of implantation is to create "channel-stop" regions between transistors. Consider the field oxide and the S/D junctions of M_1 and M_2 in Fig. 17.6(a), assuming an interconnect line passes on top of the field oxide. Interestingly, the two n^+ regions and the FOX form a MOS transistor having a thick gate oxide and hence a large threshold voltage. Nonetheless, with a sufficiently positive potential on the interconnect line, this transistor may turn on slightly, creating a leakage path between M_1 and M_2. To resolve this issue, a channel-stop implant (also called a field implant) is performed before the field oxide deposition [Fig. 17.6(b)], thereby raising the threshold voltage of the field oxide transistor to a very large value.

Ion implantation damages the silicon lattice extensively. For this reason, the wafer is subsequently heated to approximately 1000°C for 15 to 30 minutes, allowing the lattice bonds to form again. Called "annealing," this operation also leads to diffusion of dopants, broadening the profile in all directions. For example, annealing results in side-diffusion of S/D regions, creating overlap with the gate area. The wafer is therefore usually annealed only once, after all implantations have been completed.

An interesting phenomenon in ion implantation is "channeling." As shown in Fig. 17.7(a), if the implant beam is aligned with the crystal axis, the ions penetrate the wafer to a great depth. For this reason, the implant (or the wafer) is tilted by 7–9° [Fig. 17.7(b)], avoiding such an alignment and ensuring a predictable profile. As explained in Chapter 18, this tilt impacts the matching of transistors, necessitating precautions in the layout.

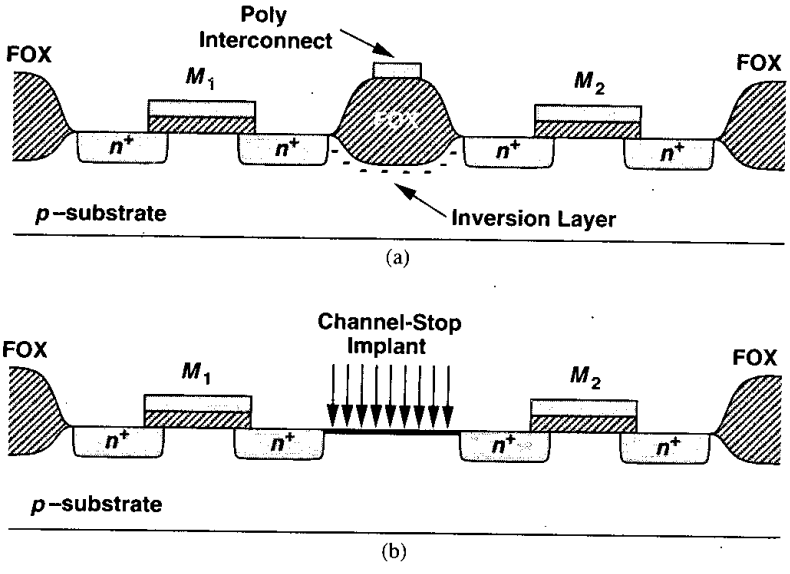


Figure 17.6 (a) Unwanted conduction due to inversion of field area, (b) channel-stop implant.

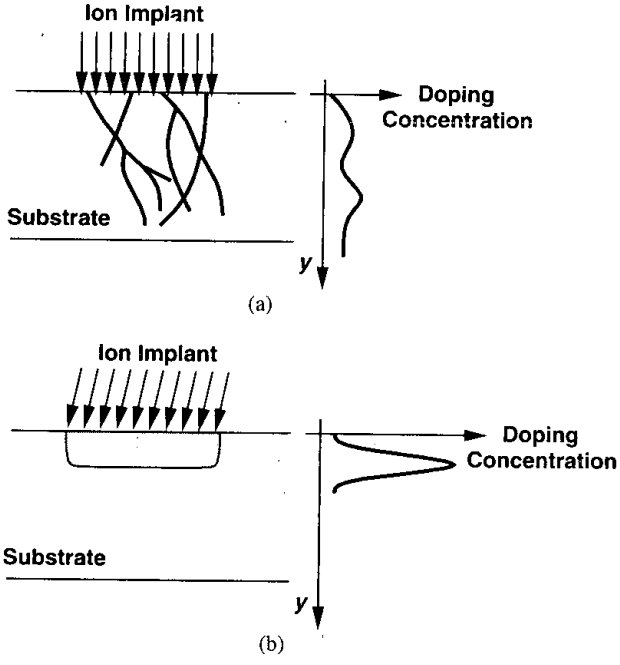


Figure 17.7 (a) Effect of channeling, (b) tilt in implant to avoid channeling.

17.6 Deposition and Etching

As suggested by the structures of Fig. 17.1, device fabrication requires the deposition of various materials. Examples include polysilicon, dielectric materials separating interconnect layers, and metal layers serving as interconnects.

A common method of forming polysilicon on thick dielectric layers is "chemical vapor deposition" (CVD), whereby wafers are placed in a furnace filled with a gas that creates the desired material through a chemical reaction. In modern processes, CVD is performed at a low pressure to achieve more uniformity.

The etching of the materials is also a crucial step. For example, contact windows with very small dimensions, e.g., 0.3 μm $\times$ 0.3 μm, and relatively large depths, e.g., 2 μm, must be etched with high precision. Depending on the speed, accuracy, and selectivity required in the etching step, and the type of material to be etched, one of these methods may be used: (1) "wet" etching, i.e., placing the wafer in a chemical liquid (low precision); (2) "plasma" etching, i.e., bombarding the wafer with a plasma gas (high precision); (3) reactive ion etching (RIE), where ions produced in a gas bombard the wafer.

17.7 Device Fabrication

With the processing operations described in the previous section, we now study the fabrication sequence and device structures in typical CMOS technologies. We consider three categories: active devices, passive devices, and interconnects.

17.7.1 Active Devices

Basic Transistor Fabrication The fabrication begins with a p-type silicon wafer approximately 1 mm thick. Following cleaning and polishing steps, a thin layer of silicon dioxide is grown as a protective coating on top of the wafer [Fig. 17.8(a)]. Next, to create the n-wells, a lithography sequence consisting of photoresist deposition, exposure to UV light using the n-well mask, and selective etching is carried out and the n-wells are implanted [Fig. 17.8(b)]. The remaining photoresist and oxide layers are then removed [Fig. 17.8(c)].

Recall from the previous section that a field implant and a field oxide growth are necessary in the areas between the transistors. At this point in the sequence, a stack consisting of a silicon oxide layer, a silicon nitride (Si_3N_4), and a *positive* photoresist layer is created. Next, the "active" mask is used for lithography so that only the regions between the transistors are exposed [Fig. 17.8(d)].[2] Subsequently, the channel-stop implant is performed, the photoresist is removed, and a thick oxide layer is grown in the exposed silicon areas, producing the field oxide. The protective nitride and oxide layers are then removed [Fig. 17.8(e)], thereby exposing all areas where transistors are to be formed. In the subsequent diagrams, the channel-stop implant will be omitted for the sake of clarity.

The next step involves the growth of the gate oxide, a critical operation requiring slow, low-pressure CVD [Fig. 17.8(f)]. As explained in Chapter 2, the "native" threshold voltage

[2]The n-wells are not shown for clarity.

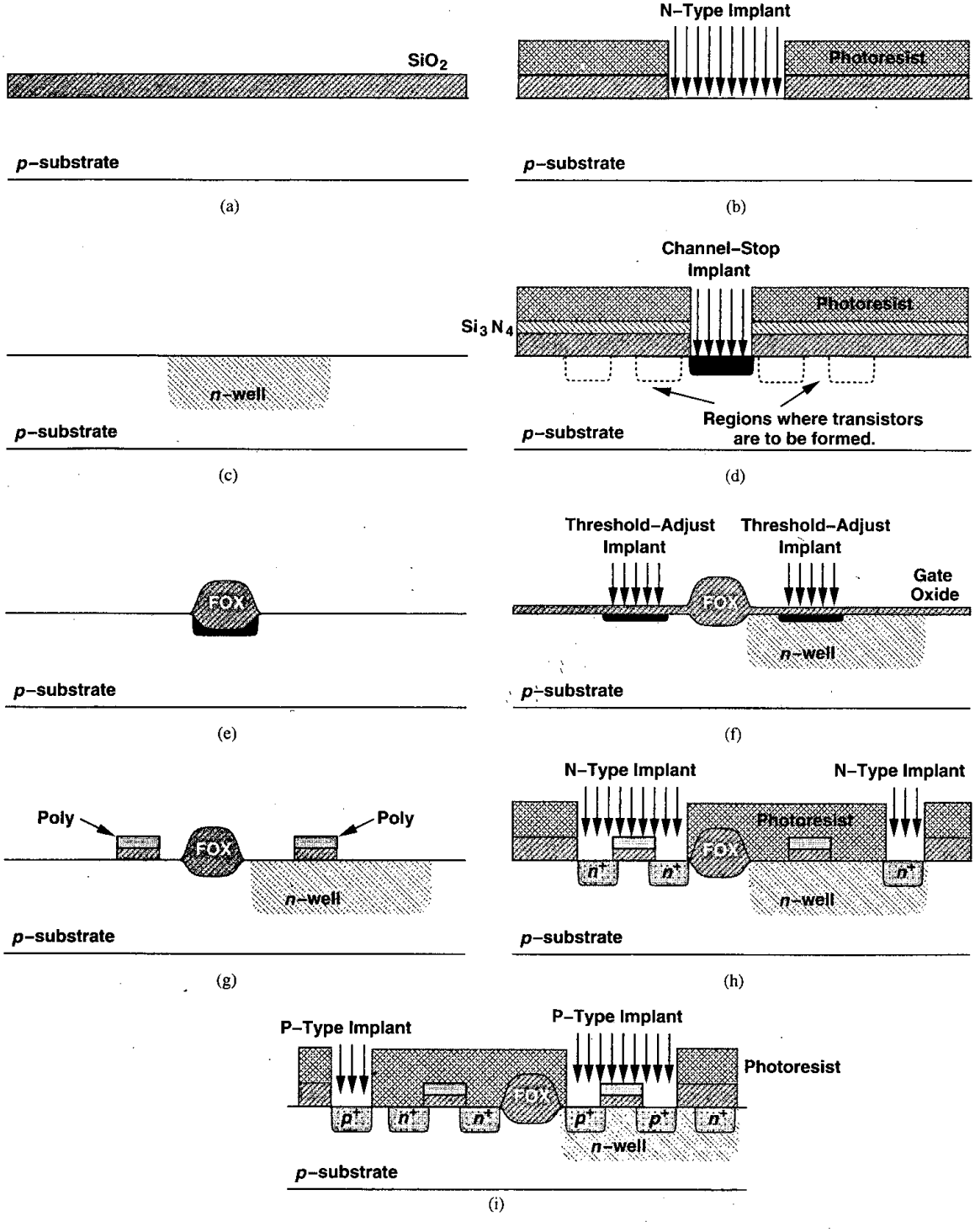


Figure 17.8 Fabrication sequence of MOS devices.

of the transistors is typically far from the desired value, necessitating a threshold-adjust implant. (The native threshold of both PMOS and NMOS is usually more negative than desired, e.g., $V_{THN} \approx 0$, and $V_{THP} \approx -1$ V.) Such an implant is performed following the growth of the gate oxide, creating a thin sheet of dopants near the surface and making the threshold of both NMOS and PMOS devices more positive.

With the gate oxide in place, the polysilicon layer is deposited and the "poly mask" lithography is carried out, resulting in the structure shown in Fig. 17.8(g). We should note that polysilicon is simply noncrystalline ("amorphous") silicon, a property that arises because this layer grows on top of silicon dioxide and hence cannot form a crystal. Since polysilicon serves as a conductor, its amorphous nature is unimportant. To reduce the resistivity of this layer, an additional implant is typically used, yielding a sheet resistance of a few tens of ohms per square.

In the next step, the source/drain junctions of the transistors and the substrate and n-well ties are formed by ion implantation. This step requires a "source/drain mask" and two lithography sequences. As illustrated in Fig. 17.8(h), the first sequence incorporates a negative photoresist, exposing the areas to receive an n^+ implant (the S/D junctions of NMOS transistors and the n-well ties). In the second sequence [Fig. 17.8(i)], the same mask and a positive photoresist are used, exposing the areas to receive a p^+ implant (the S/D junctions of PMOS transistors and the substrate ties). Note that these implants also dope the polysilicon layer, reducing its sheet resistance. This step completes the fabrication of the basic transistors.

The reader may wonder why the source/drain junctions are formed *after* the gate oxide and polysilicon. Suppose, as depicted in Fig. 17.9(a), these junctions are created first. Then, the alignment of the gate poly mask with respect to the S/D areas becomes extremely critical. Even if the misalignment is a small fraction of the minimum channel length, a gap may appear between the source (or drain) and the gate area, prohibiting the formation

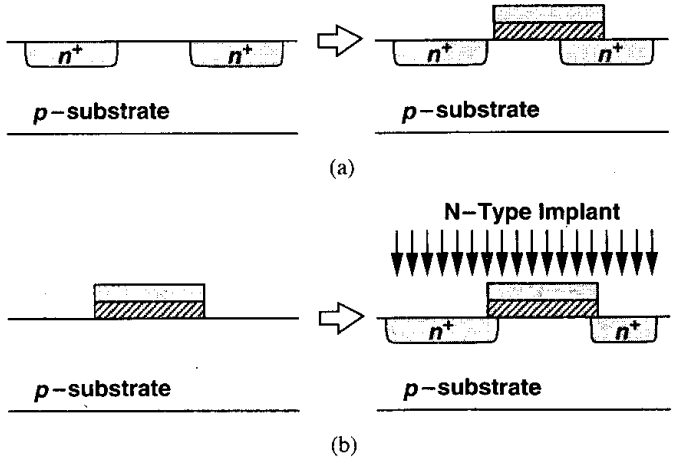


Figure 17.9 (a) Formation of n^+ regions before deposition of poly, (b) self-aligned structure.

of a continuous channel in the transistor. By contrast, the sequence shown in Fig. 17.8 yields a "self-aligned" structure because the source/drain regions are implanted at precisely the edges of the gate area and a misalignment in lithography simply makes one junction slightly narrower than the other [Fig. 17.9(b)]. Interestingly, the first few generations of CMOS technology were based on the approach shown in Fig. 17.9(a), but it was soon discovered that the self-aligned structure would lend itself to scaling much more easily.

Back-End Processing With the basic transistors fabricated, the wafers must next undergo "back-end" processing, a sequence primarily providing various electrical connections on the chip through contacts and wires. The first step in this sequence is "silicidation." Since the sheet resistance of doped polysilicon and S/D regions is typically several tens of ohms per square, it is desirable to reduce their resistance by about an order of magnitude. Silicidation accomplishes this by covering the polysilicon layer and active areas (S/D regions and substrate and n-well ties) with a thin layer of a highly conductive material, e.g., titanium silicide or tungsten. Illustrated in Fig. 17.10, this step in fact begins with creating an "oxide spacer" at the edges of the polysilicon gate such that the deposition of the silicide becomes a self-aligned process as well.[3] Without the spacer, the silicide layer on the gate may be shorted to that on the source/drain.

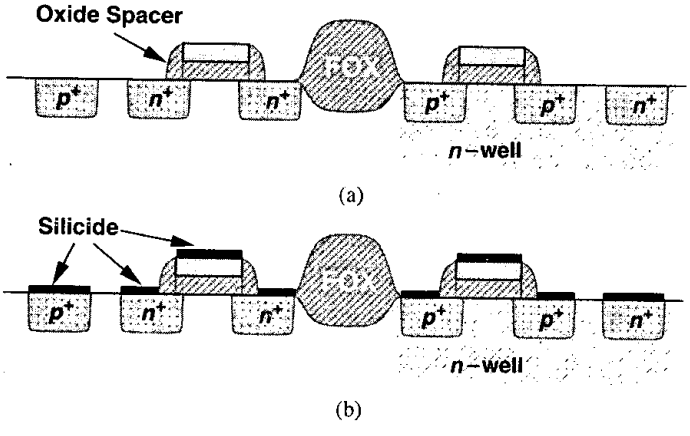


Figure 17.10 (a) Oxide spacers and (b) silicide.

The next step in back-end processing is to produce contact windows on top of polysilicon and active regions. This is carried out by first covering the wafer with a relatively thick (0.3- to 0.5-μm) layer of oxide and subsequently performing a lithography sequence using the "contact mask." The contact holes are then created by plasma etching [Fig. 17.11(a)]. Owing to reliability issues, contacts to the gate polysilicon are not placed on top of the gate area.

Following contact windows, the first layer of metal interconnect (called "metal 1") (using aluminum or copper) is deposited over the entire wafer. A lithography sequence

[3]Self-aligned silicide is sometimes called "salicide."

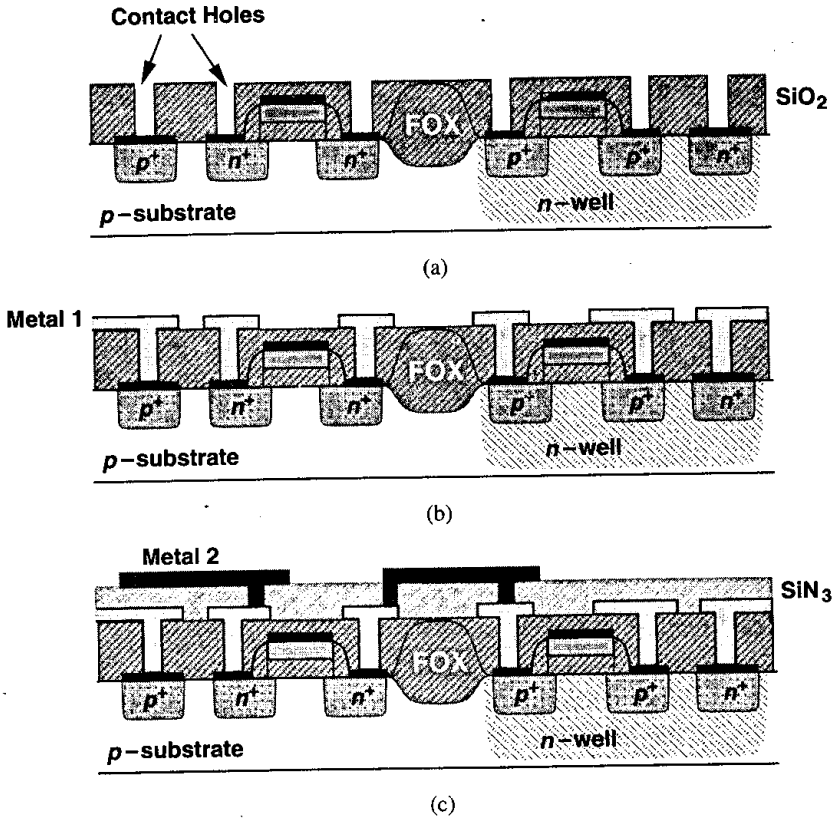


Figure 17.11 Contact and metal fabrication.

using the "metal 1 mask" is then carried out and the metal layer is selectively etched [Fig. 17.11(b)].

The higher levels of interconnect are fabricated using the same procedure [Fig. 17.11(c)]. For each additional metal layer, two masks are required: one for the contact windows and another for the metal itself. Thus, a CMOS process having five layers of metal contains 10 masks for the back end. The contact windows between metal layers are sometimes called "vias" to distinguish them from the first level of contacts to active areas and polysilicon.

We should mention that if a large area must be contacted, many small windows—rather than a large window—are usually used. Dictated by reliability issues, the dimensions of each contact or via are fixed and cannot be decreased or increased by the layout designer. An interesting phenomenon related to large active areas is "contact spiking." If a large contact window allows aluminum to touch the active area, then, as depicted in Fig. 17.12(a), the metal may "eat" and penetrate the doped region, eventually crossing the junction to the bulk and shorting the diode. With small windows, on the other hand, this effect is avoided [Fig. 17.12(b)].

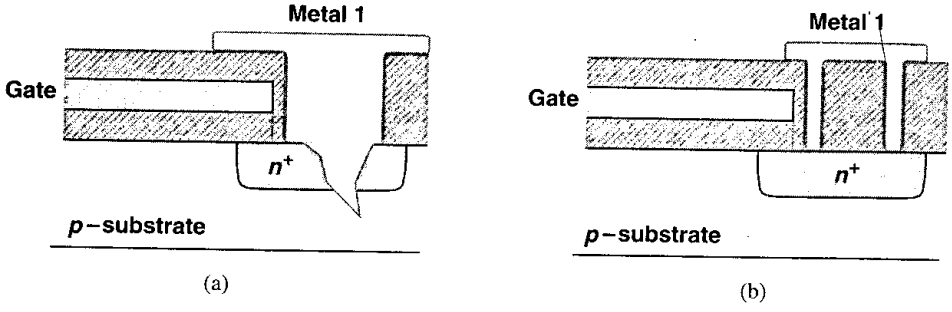


Figure 17.12 (a) Spiking due to large contact areas, (b) use of small contacts to avoid spiking.

The final step in back-end processing is to cover the wafer with a "glass" or "passivation" layer, protecting the surface against damages caused by subsequent mechanical handling and dicing. After a lithography sequence using the "passivation mask," the glass is opened only on top of the bond pads to allow connection to the external environment (e.g., the package).

17.7.2 Passive Devices

Passive components such as resistors and capacitors find wide usage in analog design, making it desirable to add these devices to standard CMOS technologies. In practice, however, CMOS processes target primarily digital applications and hence provide only NMOS and PMOS transistors. A new generation of CMOS technology may take one to two years and many iterations before it becomes an "analog process," i.e., one offering high-quality passive devices. If a digital CMOS process is to be used for analog design, we must seek structures that can serve as passive components. The principal issue in using such structures is the *variability* of the component value from wafer to wafer because the process flow does not assume such structures are used in circuits.

Resistors A CMOS process may be modified so as to provide resistors suited to analog design. A common method is to selectively "block" the silicide layer that is deposited on top of the polysilicon, thereby creating a region having the resistivity of the doped polysilicon (Fig. 17.13). This means the fabrication requires an additional mask and a corresponding lithography sequence. Since the poly doping level is determined by various implants in the process, the resistivity obtained here is not necessarily a target value, but it usually falls in the range of fifty to a few hundred ohms per square. For the same reason, the resistance value may vary by as much as $\pm 20\%$ from wafer to wafer or lot to lot.

The use of silicide on the two ends of the resistor in Fig. 17.13 results in a much lower contact resistance than that obtained by directly connecting the metal layer to doped polysilicon. This improves both the definition of the resistor value and the matching with identical structures. Also, for a given resistance, poly resistors typically exhibit much less capacitance to the substrate than other types—on the order of 90 af/μm^2 for the bottom plate capacitance and 100 af/μm for the fringing capacitance. These resistors are quite linear,

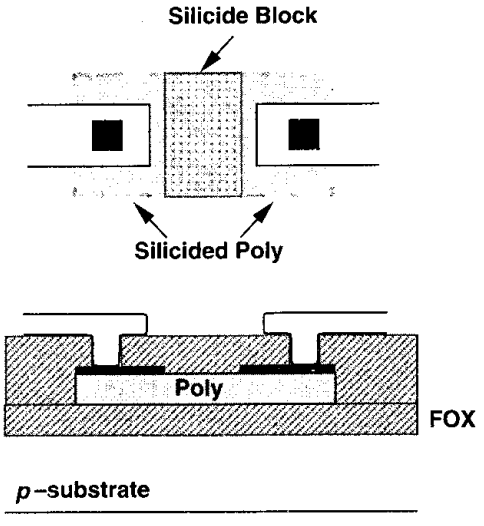


Figure 17.13 Poly resistor using silicide block.

especially if they are long. The primary difficulties with silicide-block poly resistors are variability, mask cost, and process complexity.

In a purely digital process, silicided poly, silicided p^+ or n^+ active areas, n-well, and metal layers can be used as resistors. Since the silicided layers have a low resistivity and, more importantly, their resistance varies substantially (e.g., by $\pm 50\%$), they are rarely used in analog circuits.[4] An n-well resistor can be formed as shown in Fig. 17.14, but the n-well resistivity may vary by several tens of percent with process. With typical sheet resistivities of about 1 k$\Omega/\square$, n-well resistors can prove useful where their absolute value is

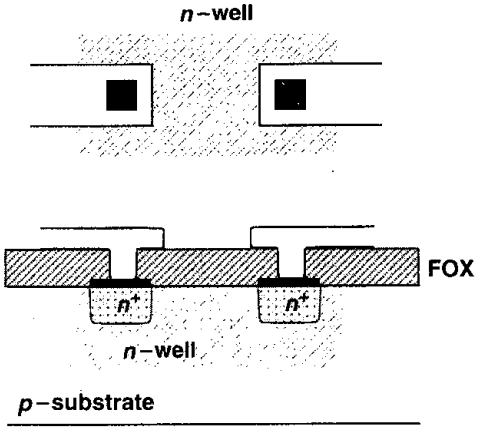


Figure 17.14 Resistor made of n-well.

[4] One exception is where the low value is desirable and the absolute value is not critical, e.g., in resistor ladders used in A/D converters (Chapter 18).

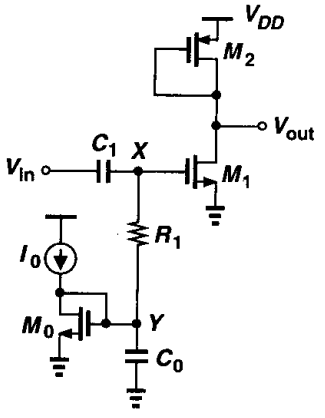


Figure 17.15 Use of an n-well resistor in a coupling network.

not critical. For example, Fig. 17.15 shows a common-source stage that is biased by means of M_0 and I_0 while employing C_1 to block the dc level of the preceding stage. In order to isolate the signal path from the low impedance (and the noise) introduced by M_0, resistor R_1 is inserted between X and Y. Here, the value of R_1 is not critical so long as it is sufficiently large.

We should mention that, due to the depletion region formed between the n-well and the p-substrate, n-well resistors suffer from both a large parasitic capacitance and significant voltage dependence. Fig. 17.16 illustrates a typical case, where one terminal of the n-well resistor is tied to V_{DD}. Since the capacitance to the substrate is distributed (nonuniformly) along the resistor, a lumped model may not be accurate enough, but as a rough approximation, we place half of the total capacitance on each side of the resistor. We also note that as V_{out} varies, so do the width of the depletion region and hence the value of the resistor.

The metal layers available in CMOS technologies exhibit sheet resistances on the order of 70 mΩ/□ (for bottom layers) to 30 mΩ/□ (for top layers). Thus, for resistor values common in analog design, metal layers are rarely used.

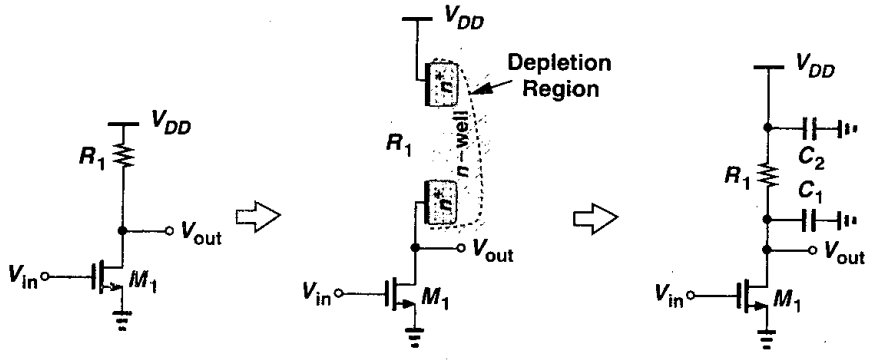


Figure 17.16 Common-source stage using n-well resistors.

Capacitors Capacitors prove indispensible in most of today's analog CMOS circuits. Several parameters of capacitors are critical in analog design: nonlinearity (voltage dependence), parasitic capacitance to the substrate, series resistance, and capacitance per unit area (density). In CMOS technologies modified for analog design, capacitors are fabricated as "poly-diffusion," "poly-poly," or "metal-poly" structures. Illustrated in Fig. 17.17, the idea is to grow or deposit a relatively thin oxide between two floating conductive layers, thereby forming a dense capacitor with moderate bottom-plate parasitic (about 10 to 20%).

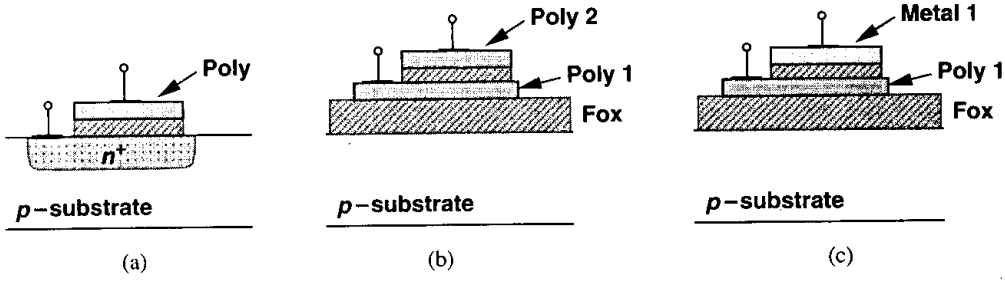


Figure 17.17 Linear capacitor structures, (a) poly-diffusion, (b) poly-poly, (c) metal-poly.

The fabrication steps required to build the poly-diffusion capacitor of Fig. 17.17(a) are shown in Fig. 17.18. First, using the "capacitor mask," a lithography sequence similar to that of Fig. 17.3 defines the bottom plate areas and a heavy n^+ implant is applied [Fig. 17.18(a)]. Next, the gate oxide layer is grown over the entire wafer [Fig. 17.18(b)]. Note that the oxide grows faster over the n^+ region because of the heavier doping level, yielding a capacitance per unit area less than that of MOSFETs. The fabrication then proceeds as

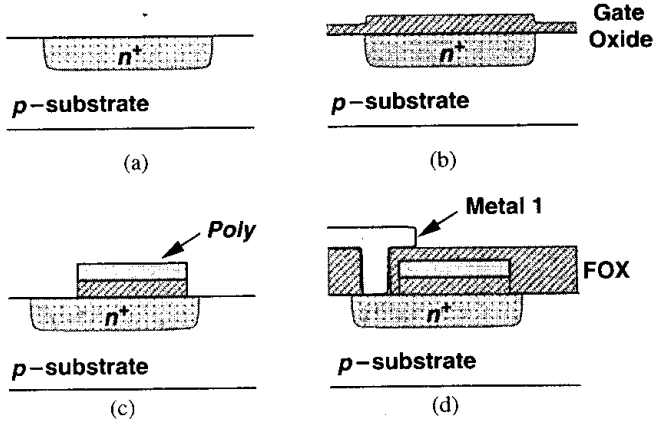


Figure 17.18 Fabrication steps of poly-diffusion capacitor.

in a standard CMOS process, forming capacitors and MOS devices simultaneously [Figs. 17.18(c) and (d)].

It is important to understand the necessity of the capacitor mask in the above sequence. The self-aligned process used to fabricate MOS devices forms active regions in the substrate only after the gate oxide and polysilicon are created. It is therefore impossible to build a doped area *under* the polysilicon without an extra lithography sequence. Even if the layout is drawn as in Fig. 17.19, since the n^+ step is performed after the deposition of poly, the result is still a MOS device rather than a linear capacitor.

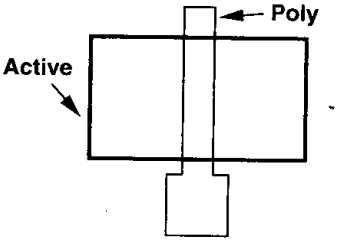


Figure 17.19 Layout yielding a MOSFET rather than a linear capacitor.

By virtue of its simplicity, the poly-diffusion capacitor is the most common type, but it still requires an additional mask and associated fabrication steps. This structure suffers from some nonlinearity because the width of the depletion regions at the poly-oxide and oxide-diffusion interfaces changes with the applied voltage (Fig. 17.20), thereby varying the equivalent dielectric thickness between the two conductive plates. If $C \approx C_0(1 + \alpha_1 V + \alpha V^2)$, then α_1 and α_2 are typically on the order of $5 \times 10^{-4}\ V^{-1}$ and $5 \times 10^{-5}\ V^{-2}$, respectively. The bottom-plate parasitic of this topology is about 20% of the interplate capacitance.

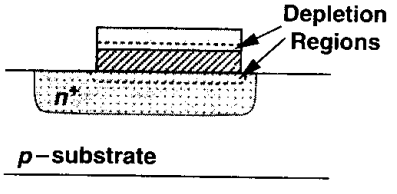


Figure 17.20 Depletion regions in a poly-diffusion capacitor.

The poly-poly capacitor of Fig. 17.17(b) is used in "double-poly" processes, e.g., those designed for fabricating electrically erasable programmable read-only memories (EEPROMs). Requiring both a capacitor mask and processing steps for the deposition and etching of the second polysilicon layer, this structure is available in some technologies and has roughly the same linearity and bottom-plate parasitic as the poly-diffusion capacitor.

The metal-poly topology shown in Fig. 17.17(c) is the most linear and the most expensive of the three. Here, after the transistors are formed and the polysilicon is silicided, a thin layer of SiO_2 is deposited over the entire wafer. Next, a lithography sequence using the capacitor mask defines areas on top of polysilicon where the oxide must remain, and selective etching

is performed. Owing to silicidation, no depletion region is formed at the poly-oxide interface and the linearity of the capacitor is improved. Nonlinearity coefficients as low as a few parts per million have been achieved for such a structure [1]. The bottom-plate parasitic is on the order of 10 to 20%.

Digital CMOS technologies do not offer the foregoing capacitor structures for cost and yield reasons. The designer must therefore construct capacitors through the use of the "native" layers of the process.

Perhaps the simplest capacitor structure in CMOS technology is that implemented by a MOSFET. Illustrated in Fig. 17.21(a), the device has a capacitance that varies from a small value at low voltages (where no channel exists and the equivalent capacitance is the series combination of the oxide capacitance and the depletion region capacitance) to a large value (C_{ox}) if the voltage difference exceeds V_{TH}. Since the gate oxide is typically the thinnest layer in the process, MOS capacitors biased in strong inversion are quite dense, saving substantial area if large values are required. For the same reason, the bottom-plate parasitic, i.e., that due to drain and source junctions, is a relatively small percentage of the gate capacitance—typically 10 to 20%.

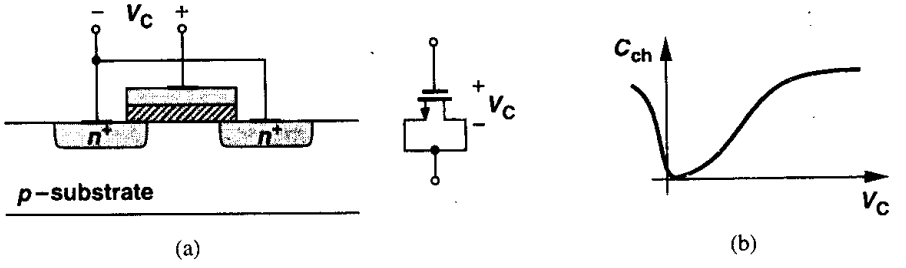


Figure 17.21 (a) MOSFET configured as a capacitor, (b) nonlinear C/V characteristic.

Unfortunately, the voltage dependence of MOS capacitors, even in strong inversion, makes the structure less attractive for precision charge transfer.

Example 17.1

Consider the multiply-by-two amplifier of Section 12.3.3, shown in Fig. 17.22(a) as an implementation using a MOS capacitor C_1 and a linear capacitor C_2. Explain how the output voltage in the amplification mode is distorted.

Solution

Suppose for simplicity that V_{in} is below ground by more than V_{TH} so that the NMOS capacitors are in strong inversion during sampling. As the circuit enters the amplification mode, the voltage across C_1 approaches zero and the total charge stored on C_1 is transferred to C_2. How much is this charge? If C_1 were linear, we would have $Q = C_1 V$, but here we must write $dQ = C_1 dV$. Thus, as shown in Fig. 17.22(b), the total transferred charge when the voltage across the capacitor goes from V_{in} to

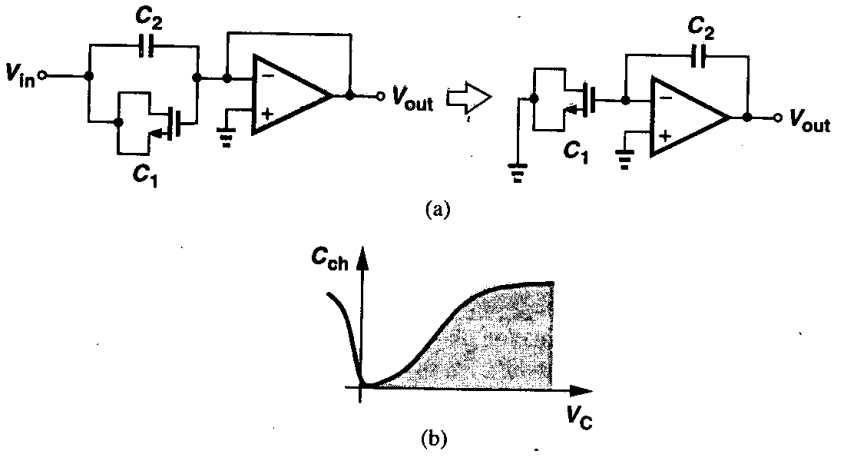


Figure 17.22 Precision multiply-by-two circuit using a MOS capacitor.

zero is equal to the area under the C/V characteristic, a value substantially less than that in the linear case. The output voltage is then given by

$$V_{out} \approx V_{in} + \frac{1}{C_2}\int_0^{V_{in}} C_1\, dV. \tag{17.1}$$

Another issue related to MOS capacitors is their series resistance, an effect arising from the gate material and, more importantly, the channel resistance. Assuming proper layout minimizes the gate resistance, we view the channel resistance as shown in Fig. 17.23, estimating the equivalent series resistance as $(R_{tot}/2)\|(R_{tot}/2) = R_{tot}/4$, where $R_{tot} = [\mu C_{ox}(W/L)(V_{GS} - V_{TH})]^{-1}$. The intrinsic time constant of the capacitor is therefore

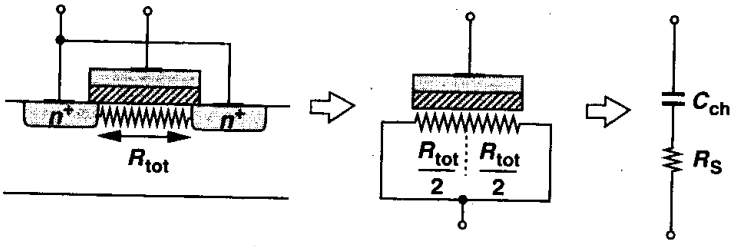


Figure 17.23 Channel resistance of MOS capacitor.

equal to:

$$\tau = \frac{R_{tot}}{4} C_{ch} \tag{17.2}$$

$$= \frac{1}{4\mu C_{ox}(W/L)(V_{GS} - V_{TH})} \cdot WLC_{ox} \tag{17.3}$$

$$= \frac{L^2}{4\mu(V_{GS} - V_{TH})}. \tag{17.4}$$

In reality, the distributed nature of the resistance and the capacitance along the channel results in a time constant equal to one-third of that given above [2]. Another figure of merit for such a capacitor is $Q = [1/(C\omega)]/R_S$. As a rule of thumb, we choose $R_S < 0.1/(C\omega)$.

Equation 17.4 indicates that for a given overdrive, to minimize the series resistance of a MOS capacitor, L must be minimized. Consequently, MOS capacitors are usually designed as a parallel combination of wide, short devices rather than a *square* block (Fig. 17.24). The penalty is a higher junction capacitance to the substrate and somewhat greater area.

In applications requiring linear capacitors, a "sandwich" of conductive layers can be formed in CMOS technology. Shown in Fig. 17.25 is an example, where the capacitance between every two layers is exploited to increase the density. Since the dielectrics between

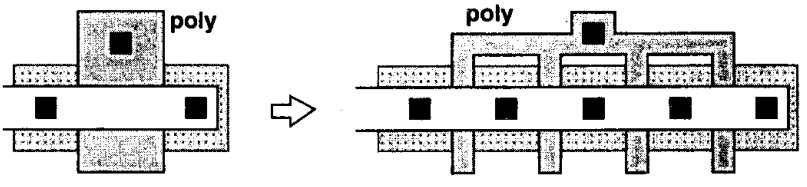


Figure 17.24 Use of wide, short MOS fingers to reduce channel resistance.

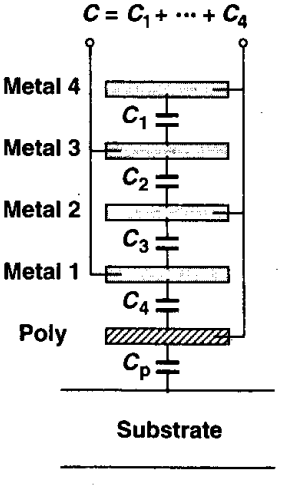


Figure 17.25 Linear capacitor made of native conductive layers.

the layers are relatively thick, this structure still requires a much larger area than the types studied above. More importantly, the bottom-plate parasitic (e.g., the capacitance between poly and substrate in Fig. 17.25) is quite large, about 50 to 60% of the total interplate capacitance. This structure is studied in detail in Chapter 18.

Example 17.2

An amplifier with an input capacitance of C_{in} is to be ac-coupled to a preceding stage having an output resistance R_{out}. Considering both of the topologies depicted in Fig. 17.26 and allowing a maximum signal attenuation of 20%, determine the minimum value of the coupling capacitor and the resulting time constant if $C_P = 0.5C_C$ or $C_P = 0.2C_C$.

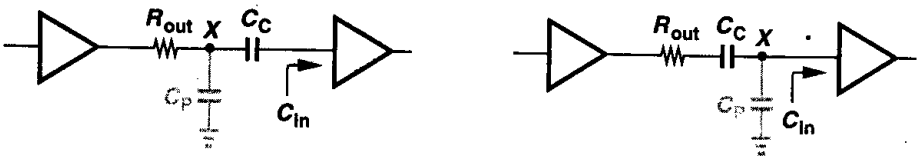


Figure 17.26

Solution

In Fig. 17.26(a), the attenuation is given by: $A_v = C_C/(C_C + C_{in})$, yielding $C_C \geq 4C_{in}$ for a 20% signal loss. The total capacitance seen from node X to ground is therefore equal to $C_P + C_C C_{in}/(C_C + C_{in}) = C_P + 0.8C_{in}$. It follows that the time constant is $2.8R_{out}C_{in}$ for $C_P = 0.5C_C$ and $1.6R_{out}C_{in}$ for $C_P = 0.2C_C$.

In Fig. 17.26(b), C_P itself attenuates the signal: $A_v = C_C/(C_C + C_{in} + C_P)$, indicating that no value of C_C can yield a signal loss of 20% if $C_P \geq 0.2C_C$.

These calculations yield two important results. First, the topology of Fig. 17.26(a) is generally preferable. Second, the addition of a coupling capacitor, e.g., to isolate the bias levels, substantially degrades the speed.

17.7.3 Interconnects

The performance of today's complex integrated circuits heavily depends on the quality of the available interconnects, requiring more metal layers in new generations of the technology.[5] Proper modeling of interconnects in a high-performance circuit is still a topic of active research, but our objective is to provide a basic understanding of the interconnect issues.

Two properties of interconnects, namely, series resistance and parallel capacitance, impact the performance, often mandating iteration between layout and circuit design. The series resistance becomes especially problematic in supply and ground lines, creating dc and transient voltage drops. Also, for long signal lines, the distributed resistance and capacitance of the wire may result in a significant delay.

[5] At the time of this writing, technologies with six layers of metal are in production.

The resistance of metal wires can be easily estimated at low frequencies, where skin effect is negligible. Typical sheet resistances are 30 mΩ/□ for the topmost (thickest) layer and 70 mΩ/□ for lower layers. The finite resistance of wires influences the choice of line widths for high-current interconnects such as supply and ground buses, as illustrated by the following example.

Example 17.3

A D/A converter incorporates N equal current sources implemented as NMOS devices each having an aspect ratio of W/L [Fig. 17.27(a)]. Assuming the interconnect between every two consecutive current sources has a small resistance, r, estimate the mismatch between I_N and I_1.

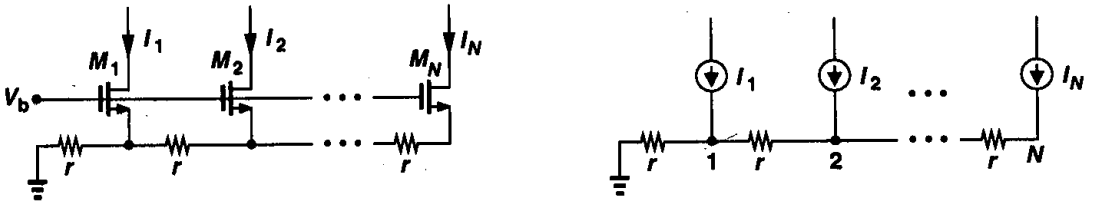


Figure 17.27 Effect of ground resistance in a D/A converter.

Solution

If r is sufficiently small, the circuit can be modeled as shown in Fig. 17.27(b), where, $I_1 \approx I_2 \approx \cdots \approx I_N = I$. The voltage at node N is obtained by superposition of currents:

$$V_N = Ir + I(2r) + \cdots + I(Nr) \tag{17.5}$$

$$= \frac{N(N+1)}{2} Ir. \tag{17.6}$$

If V_N is relatively small, the assumption $I_1 \approx I_2 \approx \cdots \approx I_N$ used in the above calculation is reasonable and M_1-M_N exhibit roughly equal transconductances. Thus,

$$I_N = I - g_m V_N \tag{17.7}$$

$$= I - g_m r \frac{N(N+1)}{2} I \tag{17.8}$$

$$= I\left[1 - g_m r \frac{N(N+1)}{2}\right]. \tag{17.9}$$

Since $V_1 \approx N \cdot I \cdot r$, we have $I_1 = I - g_m N \cdot I \cdot r$, and the relative mismatch between I_1 and I_N is

$$\left|\frac{I_1 - I_N}{I}\right| = g_m r \frac{N(N-1)}{2}. \tag{17.10}$$

The key point here is that the error grows in proportion to N^2. The ground bus must therefore be sufficiently wide to minimize r.

Another factor determining the width of interconnects is "electromigration." At high current densities, the aluminum atoms in a wire tend to "migrate," leaving a void that eventually (after some years of operation) grows to a discontinuity. For this reason, long-term reliability considerations restrict the maximum current density of interconnects. As a rule of thumb, a current-density of 1 mA per micron of width is acceptable, but the actual value varies according to the thickness of the metal. Also, for transient currents, the peak value may be quite higher.

The problem of interconnect capacitance is much more complicated. We begin with a single wire on top of a substrate (Fig. 17.28), identifying a "parallel-plate" capacitance and a "fringe" capacitance. For narrow lines, the two are comparable.

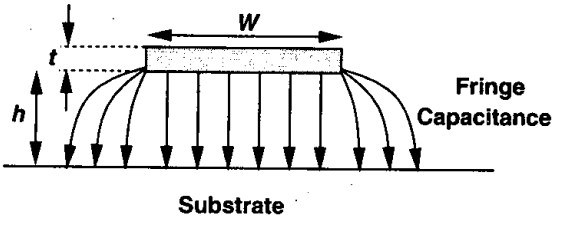


Figure 17.28 Parallel-plate and fringe capacitance of an interconnect.

A simple empirical relationship for calculating the total wire capacitance per unit length on top of a conducting substrate is:

$$C = \epsilon \left[\frac{W}{h} + 0.77 + 1.06 \left(\frac{W}{h} \right)^{0.25} + 1.06 \left(\frac{t}{h} \right)^{0.5} \right], \tag{17.11}$$

where W, h, and t denote the dimensions shown in Fig. 17.28 [3]. For typical dimensions, this equation predicts the capacitance with a few percent of error.

While upper levels of metal in a process exhibit less capacitance per unit width and length, their minimum allowable width is usually greater than that of the lower layers. Thus, the minimum capacitance for a given length may be only slightly smaller for the topmost layer(s). Table 17.1 depicts typical values for the minimum widths and parallel-plate and fringe capacitances (to the substrate) in a four-metal 0.25-μm process.

Wires also suffer from parallel and fringe capacitances between them. Illustrated in Fig. 17.29, this effect is difficult to quantify for a complex layout, often necessitating the use of computer programs. In practice, the capacitances between the layers are calculated by "electromagnetic field solvers," measured experimentally, and tabulated in the process design manual.

Table 17.1 Minimum widths and capacitances of interconnects in a 0.25-μm technology.

	Poly	Metal 1	Metal 2	Metal 3	Metal 4
Minimum Width (μm)	0.25	0.35	0.45	0.50	0.60
Bottom-Plate Capacitance (aF/μm^2)	90	30	15	9.0	7.0
Fringe Capacitance (Two Sides) (aF/μm)	110	80	50	40	30

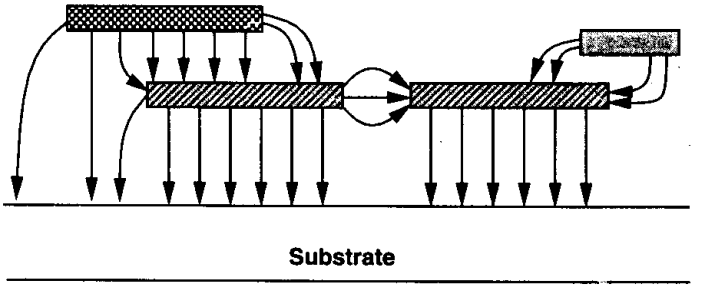


Figure 17.29 Complex interconnect structure.

17.8 Latch-Up

Owing to manufacturing difficulties, the first few generations of MOS technologies provided only NMOS devices. In fact, many of the early microprocessors and analog circuits were fabricated in NMOS processes, but they consumed substantial power. The advent of CMOS technology was motivated by the zero static power dissipation of CMOS logic—although CMOS devices required a greater number of masks and fabrication steps. Another issue that did not exist in NMOS implementations but arose in CMOS circuits was latch-up.

Consider the NMOS and PMOS devices shown in Fig. 17.30(a). Recall from Chapter 11 that a parasitic *pnp* bipolar transistor, Q_1, is associated with the PFET, the *n*-well, and the substrate. By the same token, a parasitic *npn* device, Q_2, can be identified in conjunction with the NFET. We make two observations: (1) the base of each bipolar transistor is inevitably tied to the collector of the other; (b) owing to the finite resistance of the *n*-well and the substrate, the bases of Q_1 and Q_2 see a nonzero resistance to V_{DD} and ground, respectively. The parasitic circuit can therefore be drawn as in Fig. 17.30(b), revealing a *positive* feedback loop around Q_1 and Q_2. In fact, if a current is injected into node X such that V_X rises, then I_{C2} increases, V_Y falls, $|I_{C1}|$ increases, and V_X rises further. If the loop gain is greater than or equal to unity, this phenomenon continues until both transistors turn on completely, drawing an enormous current from V_{DD}. We say the circuit is latched up.

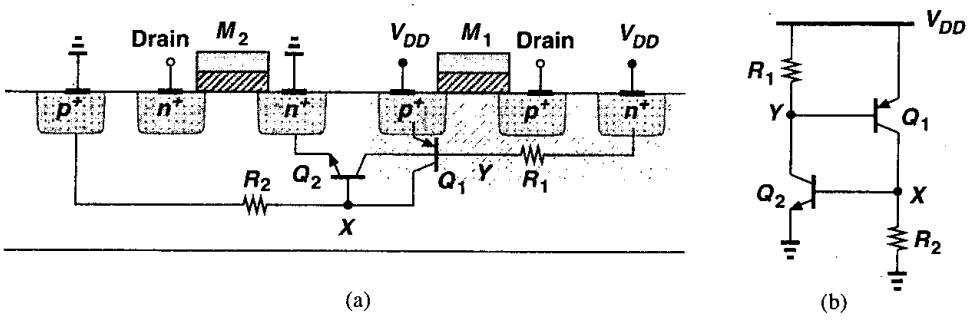


Figure 17.30 (a) Parasitic bipolar transistors in a CMOS process, (b) equivalent circuit.

The initial current required to trigger latch-up may be produced by various sources in an integrated circuit. For example, in Fig. 17.30(a), the bases of Q_1 and Q_2 are capacitively coupled to the drains of M_1 and M_2, respectively. A large voltage swing at the drains can therefore inject a significant displacement current into the n-well or the substrate, initiating latch-up.

A common case of latch-up occurs with the use of large digital output buffers (inverters). These circuits inject high currents into the substrate through the large drain junction capacitance of the transistors and by forward-biasing the source-bulk junction diodes. The latter arises because of the substantial transient voltages produced across the bond wires connected to the ground (Chapter 18).

In order to prevent latch-up, both process engineers and circuit designers take precautions such that the loop gain of the equivalent circuit shown in Fig. 17.30(b) remains well below unity. Proper choice of the doping levels and profiles as well as layout design rules ensure a low value for both the parasitic resistances and the current gain of the bipolar transistors. Furthermore, the layout of the circuit incroporates substrate and n-well contacts with sufficiently small spacing so as to minimize the resistance. The design manual of each technology typically provides an extensive set of layout rules recommended for latch-up prevention.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume $V_{DD} = 3$ V where necessary. Also, assume all transistors are in saturation.

17.1. A MOS technology is designed to provide only n-type transistors and two metal layers. Sketch the fabrication steps and determine the minimum number of masks required in this technology.

17.2. During a threshold-adjust implant, the wafer was not tilted, leading to severe channeling. Explain whether the resulting threshold voltage is higher or lower than the target value.

17.3. The circuits of Fig. 17.31 have been fabricated with a longer-than-expected gate oxidation cycle. If the threshold voltages are still equal to the desirable value, sketch V_{out} versus V_{in}

and compare the results to the target case.

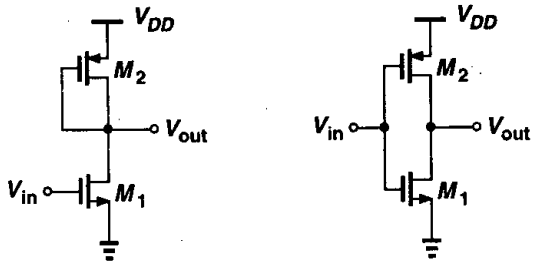


Figure 17.31

17.4. The circuits of Fig. 17.31 have been fabricated without a threshold-adjust implant. Sketch V_{out} versus V_{in} and compare the results to the target case.

17.5. Due to a layout error, the circuit shown in Fig. 17.32 suffers from contact spiking in one of the junctions. Identify the faulty junction if (a) the voltage gain is higher than expected, (b) the output voltage is near V_{DD}.

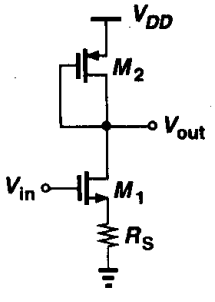


Figure 17.32

17.6. An NMOS cascode current source used in a large circuit exhibits a substantially lower output impedance than expected. Determine which fabrication error may have led to this effect: (a) channeling during S/D implant, (b) omission of the channel-stop implant, (c) insufficient gate oxide growth.

17.7. An NMOS cascode current source has a zero output current. If a single (small) lithography misalignment has caused this error, determine in which fabrication step(s) this may have occurred.

17.8. A differential pair using an active current mirror as load suffers from a low small-signal voltage gain. If the bias current is equal to the target value, determine which fabrication error may have led to this effect: (a) heavy n-well implantation, (b) heavy threshold-adjust implantation, (c) long gate oxidation cycle.

17.9. The switched-capacitor amplifier of Fig. 17.33 exhibits a large gain error. If the bias current of the op amp is equal to the desired value, which fabrication error is likely to have happened: (a) heavy threshold-adjust implantation, (b) very heavy doping in the bottom plate of C_1 (placed at node P), (c) channeling during the S/D implantation.

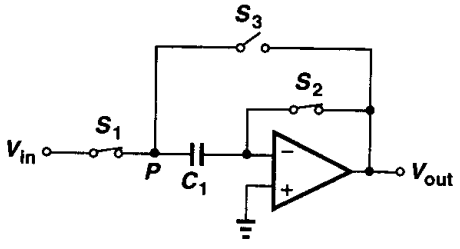


Figure 17.33

17.10. In Fig. 17.34, the digital circuit draws large transient currents from V_{DD}. Without M_1, the inductor L_b would sustain a large transient voltage $L_b dI_{DD}/dt$. Transistor M_1 with $W/L = 100/0.5$ is added to suppress this effect.

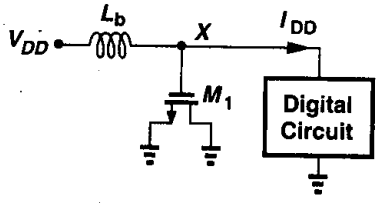


Figure 17.34

(a) Calculate the equivalent series resistance of M_1.

(b) Calculate the maximum value of L_b that results in a critically-damped response at node X. Model the digital circuit by a transient current source.

17.11. In the circuit of Fig. 17.27, $V_b = 1.2$ V, $N = 32$, and $(W/L)_{1-N} = 20/0.5$. Determine the maximum value of r for a maximum current mismatch of 1%.

17.12. Suppose in Eq. (17.11), $t = 1\ \mu$m and $h = 3\ \mu$m. For what value of W are the parallel-plate and fringe capacitances equal? What if $h = 5\ \mu$m?

References

1. C. Kaya et al., "Polycide/Metal Capacitors for High Precision A/D Converters," *IEDM Dig. of Tech. Papers,* pp. 782–785, Dec. 1988.
2. P. Larsson, "Parasitic Resistance in an MOS Transistor Used as On-Chip Decoupling Capacitor," *IEEE J. Solid-State Circuits,* vol. 32, pp. 574–576, April 1997.
3. E. Barke, "Line-to-Ground Capacitance Calculations for VLSI: A Comparison," *IEEE Trans. on Commputer-Aided Design,* vol. 7, pp. 195–298, Feb. 1988.

Chapter 18

Layout and Packaging

In the past 20 years, analog CMOS circuits have evolved from low-speed, low-complexity, small-signal, high-voltage topologies to high-speed, high-complexity, low-voltage "mixed-signal" systems containing a great deal of digital circuitry. While device scaling has enhanced the raw speed of transistors, unwanted interaction between different sections of integrated circuits as well as nonidealities in the layout and packaging increasingly limit both the speed and the precision of such systems. Today's analog circuit design is very heavily influenced by layout and packaging.

In this chapter, we study principles of layout and packaging, emphasizing effects that manifest themselves when analog and digital circuits coexist on a chip. For the sake of brevity, we use the term analog to mean both "analog" and "mixed-signal." Beginning with an overview of layout design rules, we study a number of topics related to the layout of analog circuits, including multifinger transistors, symmetry, reference distribution, passive device layout, and interconnects. Next, we deal with the problem of substrate coupling. Finally, we describe packaging issues, analyzing the effect of self- and mutual inductance and capacitance of external connections to integrated circuits.

18.1 General Layout Considerations

The layout of an integrated circuit defines the geometries that appear on the masks used in fabrication. From Chapter 17, the geometries include n-well, active, polysilicon, n^+ and p^+ implants, interlayer contact windows, and metal layers.

Figure 18.1 shows an example, where the mask geometries required for a PMOS transistor are drawn. It is important to note the following: (1) the n-well surrounds the device with enough margin to ensure that the transistor is contained in the well for all expected misalignments during fabrication; (2) each active area (S/D regions and n^+ contact to the well) is surrounded by a proper implant geometry with enough margin; (3) from the fabrication steps described in Chapter 17, the gate requires its own mask; (4) the contact windows mask provides connection from active and poly regions to the first layer of metal.

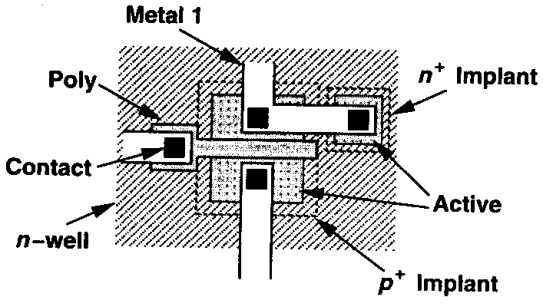


Figure 18.1 Layout of a PMOS transistor.

In most modern layout tools, the implants, and even the n-wells are automatically generated from the remainder of transistor geometries, reducing the number of layers that the layout designer sees on the computer screen and simplifying the task.

18.1.1 Design Rules

While the width and length of each transistor is determined by circuit design, most of the other dimensions in a layout are dictated by "design rules," i.e., a set of rules that guarantees proper transistor and interconnect fabrication despite various tolerances in each step of processing. Most design rules can be categorized under one of four groups described below.

Minimum Width The widths (and lengths) of the geometries defined on a mask must exceed a minimum value imposed by both lithography and processing capabilities of the technology. For example, if a polysilicon rectangle is excessively narrow, then, owing to fabrication tolerances, it may simply break or at least suffer from a large local resistance (Fig. 18.2). In general, the thicker a layer, the greater its minimum allowable width, indicating that as technologies scale, the thickness must be decreased proportionally. Fig. 18.3 depicts examples of minimum widths in a 0.25-μm technology. Note that the thickness of the layers is not under the control of the layout designer.

Figure 18.2 Excessive width variation in a narrow poly line.

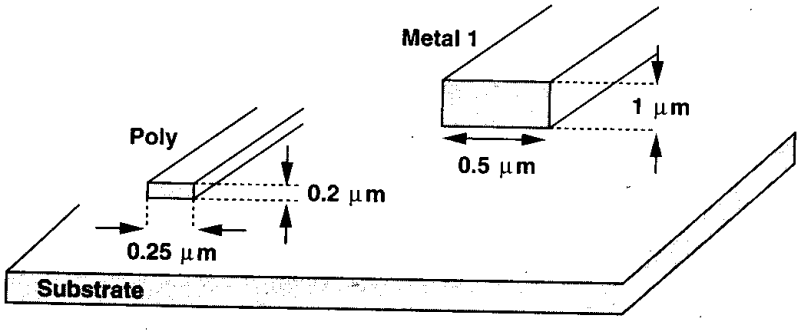


Figure 18.3 Widths and thicknesses of poly and metal lines.

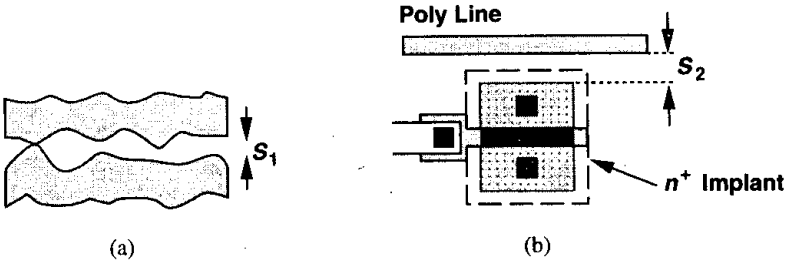


Figure 18.4 (a) Short between two excessively close poly lines, (b) minimum spacing between active and poly.

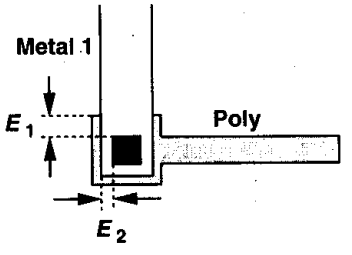


Figure 18.5 Enclosure rule for poly and metal surrounding a contact.

Minimum Spacing The geometries built on the same mask or, in some cases, different masks must be separated by a minimum spacing. For example, as shown in Fig. 18.4(a), if two polysilicon lines are placed too close to each other, they may be shorted. As another example, consider the case shown in Fig. 18.4(b), where a polysilicon line runs close to the S/D area of a transistor. A minimum spacing is required here to ensure the implant surrounding the transistor does not overlap with the poly line.

Minimum Enclosure We mentioned above that in the layout of Fig. 18.1, the n-well and the p^+ implant must surround the transistor with sufficient margin to guarantee that the device is contained by these geometries despite tolerances. These are examples of minimum enclosure rules. Fig. 18.5 depicts another example, where a poly contact window connects a poly line to a metal 1 line. To ensure that the contact remains inside the poly and metal 1 squares, both geometries must enclose the contact with enough margin.

Minimum Extension Some geometries must extend beyond the edge of others by a minimum value. For example, as shown in Fig. 18.6, the gate polysilicon must have a minimum extension beyond the active area to ensure proper transistor action at the edge.

In addition to the minimum dimensions specified in the above four categories, some *maximum allowable* dimensions may also be enforced. For example, for long metal wires, the minimum width is typically larger than that for short wires to avoid "liftoff" problems. Other such rules relate to the "antenna effect," described in the next section.

Fig. 18.7 summarizes a small subset of design rules governing the layout of an NMOS differential pair with PMOS current-source loads. Modern CMOS technologies typically involve more than 150 layout design rules.

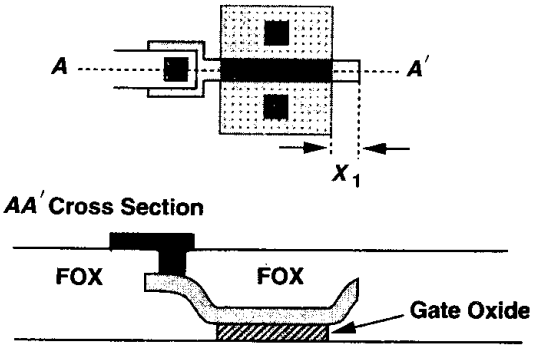
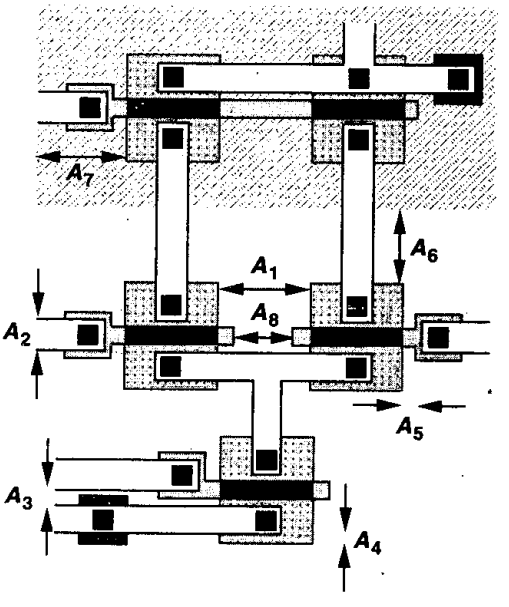


Figure 18.6 Extension of poly beyond the gate area.



A_1: Active–Active Spacing
A_2: Metal Width
A_3: Metal–Metal Spacing
A_4: Enclosure of Contact by Active
A_5: Poly–Active Spacing
A_6: Active–Well Spacing
A_7: Enclosure of Active by Well
A_8: Poly–Poly Spacing

Figure 18.7 Layout of a differential pair with PMOS current-source loads.

18.1.2 Antenna Effect

Suppose the gate of a small MOSFET is tied to a metal 1 interconnect having a large area [Fig. 18.8(a)]. During the etching of metal 1, the metal area acts as an "antenna," collecting ions and rising in potential. It is therefore possible that the gate voltage of the MOS device increases so much that the gate oxide breaks down (irreversibly) during fabrication.

The antenna effect may occur for any large piece of conductive material tied to the gate, including polysilicon itself. For this reason, submicron CMOS technologies typically limit the total area of such geometries, thereby minimizing the probability of gate oxide damage.

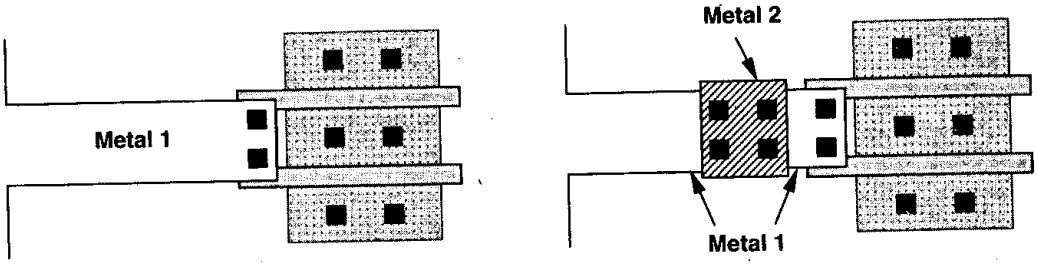


Figure 18.8 (a) Layout susceptible to antenna effect, (b) discontinuity in metal 1 layer to avoid antenna effect.

If large areas are inevitable, then a discontinuity can be created as illustrated in Fig. 18.8(b) so that, when metal 1 is being etched, the large area is not connected to the gate.

18.2 Analog Layout Techniques

The extensive sets of design rules enforced by mainstream CMOS processes aim to maximize the yield of digital ICs while allowing moderately aggressive circuit design. Analog systems, on the other hand, demand many more layout precautions so as to minimize effects such as crosstalk, mismatches, noise, etc.

18.2.1 Multifinger Transistors

As mentioned in Chapter 2, wide transistors are usually "folded" so as to reduce both the S/D junction area and the gate resistance. A simple folded structure such as that in Fig. 18.9(a) may prove inadequate for very wide devices, necessitating the use of multiple "fingers" [Fig. 18.9(b)]. As a rule of thumb, the width of each finger is chosen such that the resistance of the finger is less than the inverse transconductance associated with the finger. In low-noise applications, the gate resistance must be one-fifth to one-tenth of $1/g_m$.

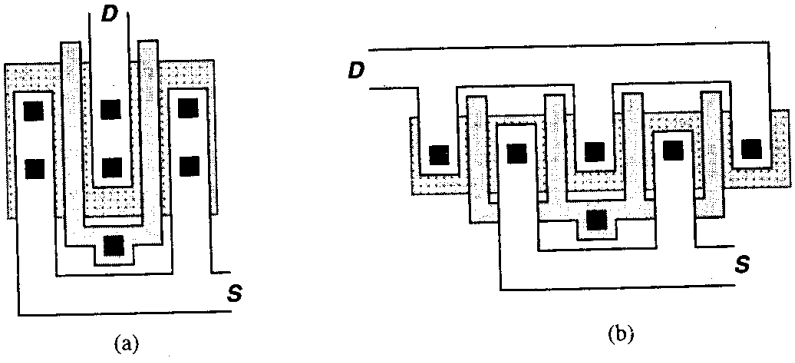


Figure 18.9 (a) Simple folding of a MOSFET, (b) use of multiple fingers.

Example 18.1

A 100-μm/0.6-μm MOSFET biased at 1 mA exhibits a transconductance of $1/(200\ \Omega)$. If the sheet resistance of the gate polysilicon is equal to 5 $\Omega/\square$, what is the widest finger that the structure can incorporate while ensuring the gate thermal noise voltage is one-fifth of the input-referred channel thermal noise voltage?

Solution

If the transistor is laid out as N parallel fingers, each finger exhibits a transconductance of $1/(200N\ \Omega)$ and a total distributed resistance of $5\ \Omega \times (100/0.6)/N$. Using the long-channel approximation for the input-referred channel thermal noise from Chapter 7, we have

$$\text{Channel Noise} = \sqrt{4kT\frac{2}{3}(200)}\ \text{V}/\sqrt{\text{Hz}} \tag{18.1}$$

$$\text{Gate Noise} = \sqrt{4kT\frac{500}{0.6N^2}\frac{1}{3}}\ \text{V}/\sqrt{\text{Hz}} \tag{18.2}$$

where the factor $1/3$ on the right hand side of (18.2) accounts for the distributed nature of the resistance (Chapter 7). Equating (18.1) to five times (18.2), we have

$$N = 5\sqrt{\frac{6.25}{3}} \tag{18.3}$$

$$\approx 7.2. \tag{18.4}$$

Thus, a minimum of 8 fingers is required.

While the gate resistance can be reduced by decomposing the transistor into more parallel fingers, the capacitance associated with the perimeter of the source/drain areas increases. As exemplified by the structures depicted in Fig. 18.10,[1] with three fingers, the total perimeter of the source or the drain is equal to $2(2E + 2W/3) = 4E + 4W/3$, whereas with five

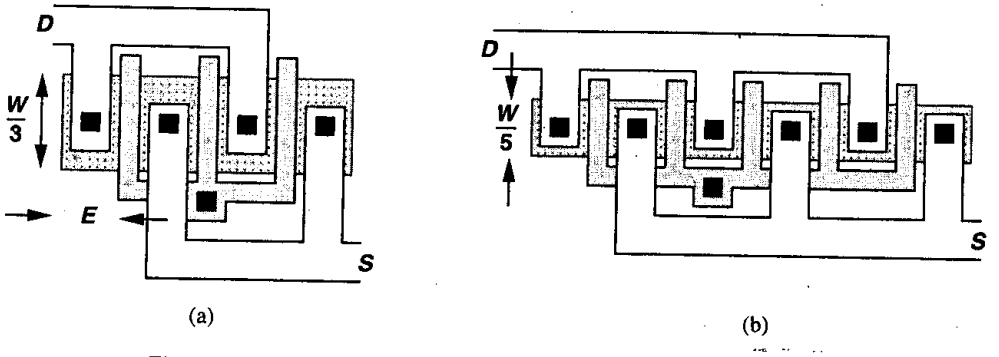


Figure 18.10 Layout of a transistor using (a) three fingers, (b) five fingers.

[1]The use of multiple fingers is sometimes called "interdigitization."

fingers, it is equal to $3(2E+2W/5) = 6E+6W/5$. In general, for an odd number of fingers N, the S/D perimeter capacitance is given by

$$C_P = \frac{N+1}{2}\left(2E + \frac{2W}{N}\right) C_{jsw} \tag{18.5}$$

$$= \left[(N+1)E + \frac{N+1}{N} W\right] C_{jsw}. \tag{18.6}$$

Thus, the number of fingers multiplied by E must be much less than W so as to minimize the S/D perimeter capacitance contribution. In practice, this requirement may conflict with that for minimizing the gate resistance noise, mandating a compromise between the two or contacting the gate on both ends to reduce the resistance.

For transistors having a large number of gate fingers, the structure may be modified to that shown in Fig. 18.11, thereby avoiding long geometries and hence disproportionate dimensions in the layout of the overall circuit.

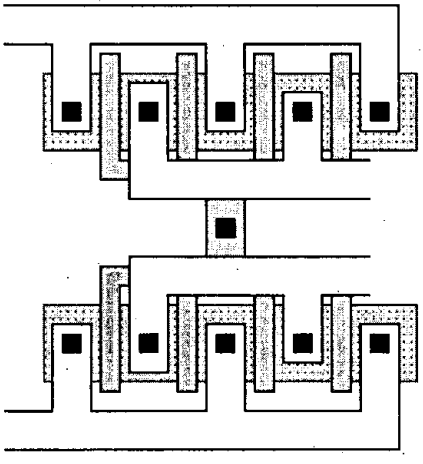


Figure 18.11 Layout of a wide transistor with many fingers.

The layout of a cascode circuit can be simplified if the input device M_1 and the cascode device M_2 have equal widths. As shown in Fig. 18.12(a), the drain of M_1 and the source of M_2 can share the same junction. More importantly, since this junction is not connected to any other node, it need not accommodate a contact window and can therefore be quite smaller [Fig. 18.12(b)]. Consequently, the capacitance at the drain of M_1 is reduced substantially, improving the high-frequency performance. For wide transistors, each transistor may use two or more fingers [Fig. 18.12(c)].

18.2.2 Symmetry

Recall from Chapter 13 that asymmetries in fully differential circuits introduce input-referred offsets, thus limiting the minimum signal level that can be detected. While some

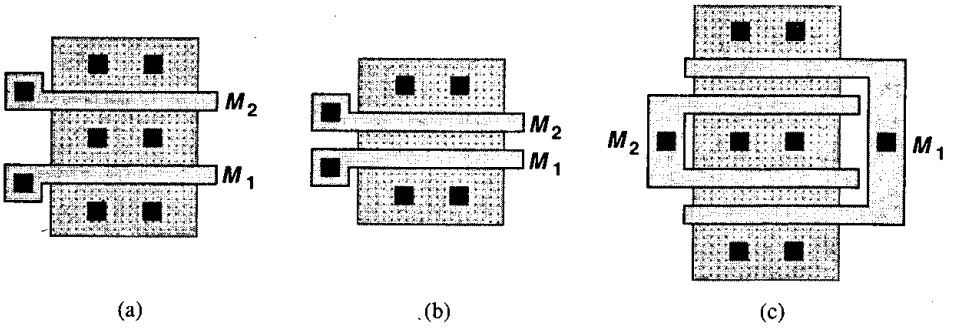


Figure 18.12 Layout of cascode devices having the same width.

mismatch is inevitable, inadequate attention to symmetry in the layout may result in large offsets—much greater than the values predicted by the statistical treatment of Chapter 13. Symmetry also suppresses the effect of common-mode noise and even-order nonlinearity. It is important to note that symmetry must be applied to both the devices of interest and their surrounding environment. We return to this point later.

Let us consider the differential pair of Fig. 18.13(a) as the starting point. If, as depicted in Fig. 18.13(b), the two transistors are laid out with different orientations, the matching greatly suffers because many steps in lithography and wafer processing behave differently along different axes. Thus, one of the configurations in Fig. 18.13(c) and (d) provides a more

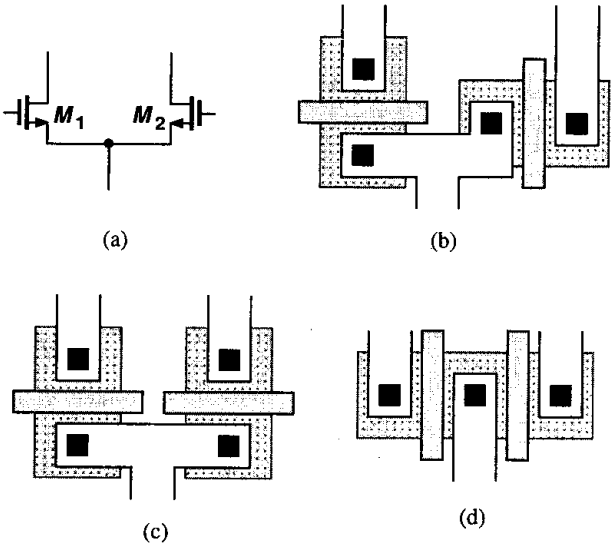


Figure 18.13 (a) Differential pair, (b) layout of M_1 and M_2 with different orientations, (c) layout with gate-aligned devices, (d) layout with parallel-gate devices.

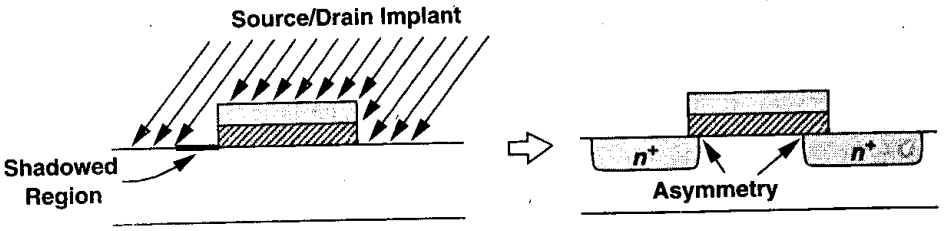


Figure 18.14 Shadowing due to implant tilt.

plausible solution. The choice between these two is determined by a subtle effect called "gate shadowing." Illustrated in Fig. 18.14, the shadowing is caused by the gate polysilicon during the source/drain implantation because the implant (or the wafer) is tilted by about 7° to avoid channeling (Chapter 17). As a result, a narrow strip in the source or drain region receives less implantation, creating a small asymmetry between the source and drain side diffusions after the implanted areas are annealed.

Now consider the structures of Figs. 18.13(c) and (d) in the presence of gate shadowing (Fig. 18.15). In Fig. 18.15(a), if the shadowed terminal is distinguished as the drain (or the source), then the two devices sustain no asymmetry resulting from shadowing. In Fig. 18.15(b), on the other hand, the transistors are not identical even if the shadowed terminals are distinguished because the source region of M_1 "sees" M_2 to its right whereas the source region of M_2 sees only the field oxide. Similarly, the drains of M_1 and M_2 see different structures to their left. In other words, the surrounding environment of M_1 is not identical to that of M_2. For this reason, the topology of Fig. 18.15(a) is preferable.

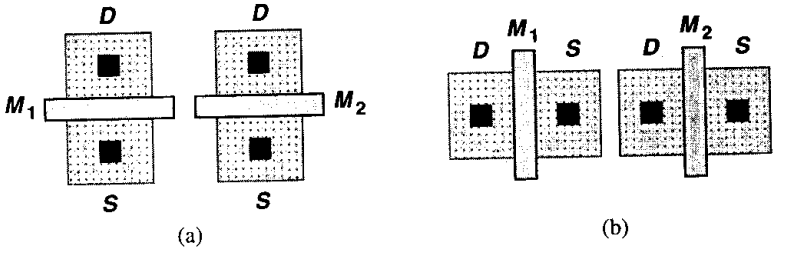


Figure 18.15 Effect of shadowing on (a) gate-aligned and (b) parallel-gate transistors.

The asymmetry inherent to the structures of Fig. 18.15(b) can be ameliorated by adding "dummy" transistors to the two sides so that M_1 and M_2 see approximately the same environment (Fig. 18.16). However, in more complex circuits, e.g., in a folded-cascode op amp, such measures cannot be easily applied.

We should emphasize the importance of maintaining the same environment on the two sides of the axis of symmetry. For example, in the structure of Fig. 18.17, an unrelated metal line passing over only one transistor indeed degrades the symmetry, increasing the mismatch between M_1 and M_2. In such cases, either a replica must be produced on the other

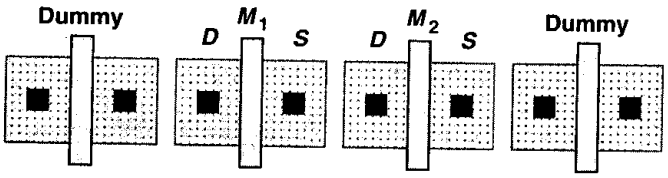


Figure 18.16 Addition of dummy devices to improve symmetry.

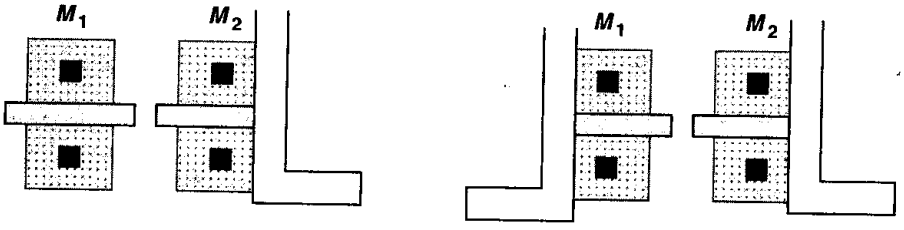


Figure 18.17 (a) Asymmetry resulting from a metal line passing over M_2, (b) removing the asymmetry by replicating the line on top of M_1.

side [Fig. 18.17(b)] (even though the replica may remain floating) or, preferably, the source of asymmetry must be removed.

Symmetry becomes more difficult to establish for large transistors. In the differential pair of Fig. 18.18, for example, the two transistors have a large width so as to achieve a small

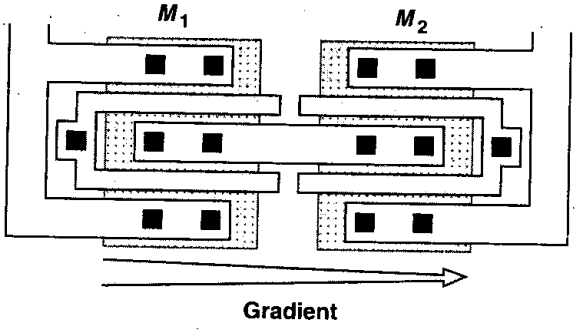


Figure 18.18 Effect of gradient in a differential pair.

input offset voltage, but gradients along the x-axis give rise to appreciable mismatches. To reduce the error, a "common-centroid" configuration may be used such that the effect of first-order gradients along both axes is cancelled. Illustrated in Fig. 18.19, the idea is to decompose each transistor into two halves that are placed diagonally opposite of each other

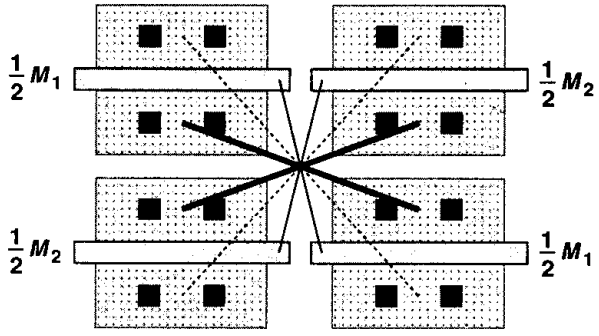


Figure 18.19 Common-centroid layout.

and connected in parallel.[2] However, the routing of interconnects in this layout is quite difficult, often leading to systematic asymmetries of the type depicted in Fig. 18.17(a) or in the capacitances from the wires to ground and between the wires. For a larger circuit, e.g., an op amp, the routing may become prohibitively complex.

The effect of linear gradients can also be suppressed by "one-dimensional" cross coupling, as depicted in Fig. 18.20. Here, all four half transistors are placed along the same axis and M_1 and M_2 are formed by connecting either the near ones and the far ones [Fig. 18.20(a)] or every other one [Fig. 18.20(b)]. (For clarity, the connections between the sources and drains are not shown.) To analyze the effect of gradients in these structures, let us assume that, for example, the gate oxide capacitance varies by ΔC_{ox} from each half transistor to the next.[3] Placing M_{1a} and M_{4a} in parallel, we have

$$I_{D1a} + I_{D4a} = \frac{1}{2}\mu_n(C_{ox} + C_{ox} + 3\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2, \tag{18.7}$$

and for M_{2a} and M_{3a}:

$$I_{D2a} + I_{D3a} = \frac{1}{2}\mu_n(C_{ox} + \Delta C_{ox} + C_{ox} + 2\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2. \tag{18.8}$$

This type of cross coupling therefore cancels the effect of the gradient. Now, for the configuration of Fig. 18.20(b), we have

$$I_{D1b} + I_{D3b} = \frac{1}{2}\mu_n(C_{ox} + C_{ox} + 2\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2, \tag{18.9}$$

and

$$I_{D2b} + I_{D4b} = \frac{1}{2}\mu_n(C_{ox} + \Delta C_{ox} + C_{ox} + 3\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2. \tag{18.10}$$

Equations (18.9) and (18.10) suggest that this approach removes the error to a lesser extent.

[2]The interconnect lines shown in this figure are only conceptually correct.

[3]In reality, variation of C_{ox} influences the threshold voltage as well. We neglect this effect here.

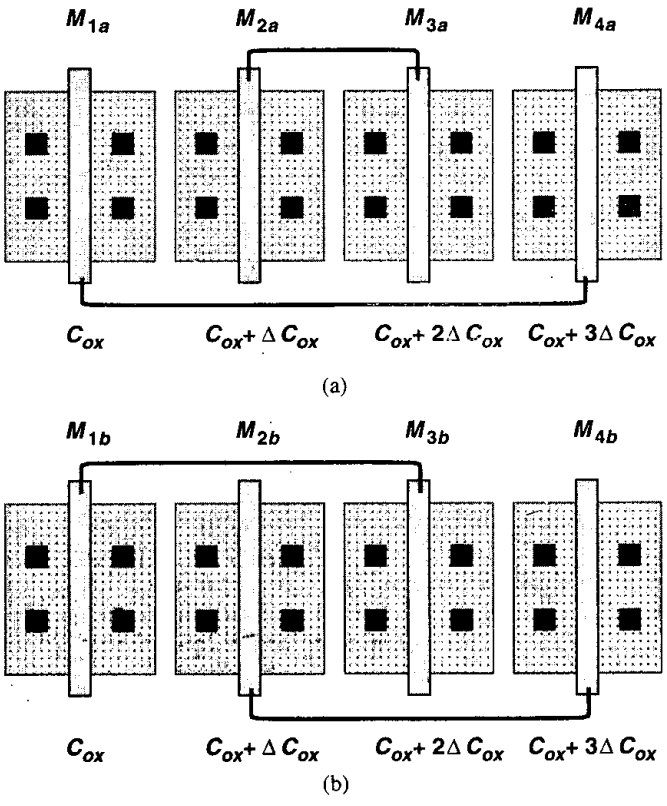


Figure 18.20 One-dimensional cross-coupling.

The reader can prove that for small gradients in other device parameters, similar results are obtained, concluding that the topology of Fig. 18.20(a) contains smaller errors than that of Fig. 18.20(b). However, since the environment seen by $M_{2a} + M_{3a}$ differs from that seen by $M_{1a} + M_{4a}$, dummy transistors must be added to the left of M_{1a} and right of M_{4a}.

18.2.3 Reference Distribution

In analog systems, the bias currents and voltages of various building blocks are derived from one or more bandgap reference generators. The distribution of such references across a large chip entails a number of important issues. Consider the example depicted in Fig. 18.21, where I_{REF} is produced by a bandgap reference and M_1-M_n serve as bias current sources of building blocks that are located far from M_{REF} and from each other. If the matching between I_{D1}-I_{Dn} and I_{REF} is critical, then the voltage drop along the ground line must be taken into account. In fact, for a large number of circuits connected to the same ground line, the systematic mismatch between the current sources and I_{REF} may be unacceptable.

To remedy the above difficulty, the reference can be distributed in the current domain rather than in the voltage domain. Illustrated in Fig. 18.22, the idea is to route the reference

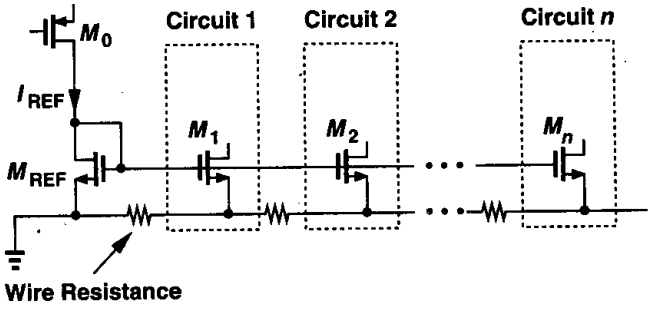


Figure 18.21 Distribution of a reference voltage for current-mirror biasing.

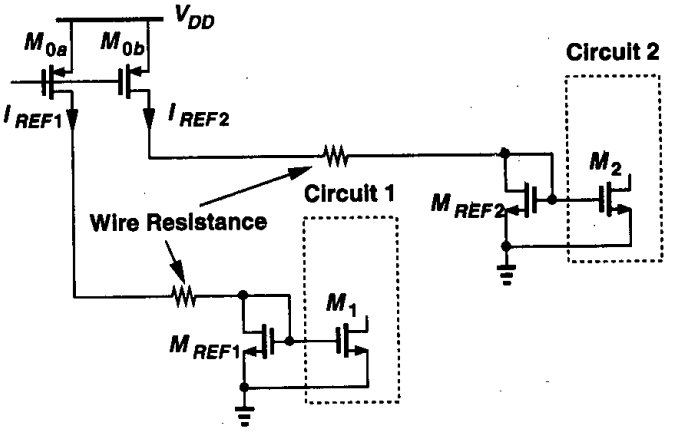


Figure 18.22 Distribution of current to reduce the effect of interconnect resistance.

current to the vicinity of the building blocks and perform the current mirror operation *locally*. Placing the interconnect resistance in series with current sources, this approach lowers systematic errors if the building blocks appear in dense groups in different regions on the chip. However, mismatches between I_{REF1} and I_{REF2} and between M_{REF1} and M_{REF2} introduce error. In large systems, it may be advantageous to employ several local bandgap reference circuits so as to alleviate routing problems.

Another issue in the circuits of Figs. 18.21 and 18.22 relates to the orientation of the transistors. As mentioned in Section 18.2.2, if, for example, M_{REF} and M_1-M_n in Fig. 18.21 have different orientations, then substantial mismatches arise. Since circuits 1, 2, . . . , n may be laid out individually, particular attention must be paid to the orientation of their current sources before and after the entire chip is composed.

The scaling of currents in Figs. 18.21 and 18.22 also demands careful choice of device dimensions and layout. Suppose the circuit of Fig. 18.21 requires $I_{D1} = 0.5I_{REF}$ and $I_{D2} =$

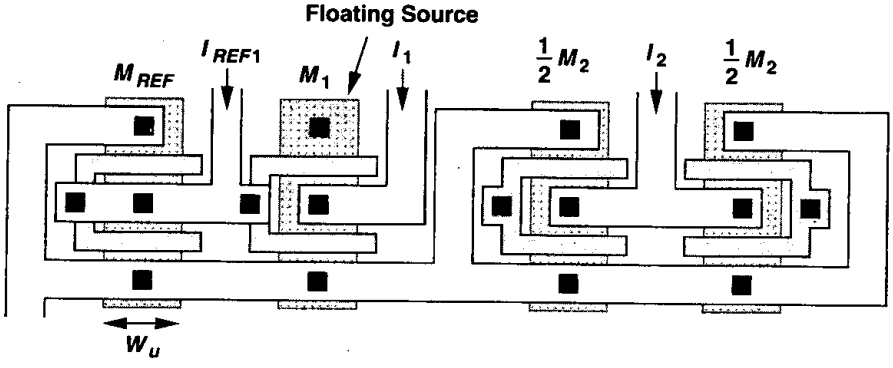


Figure 18.23 Proper scaling of device dimensions for adequate matching of current sources.

$2I_{REF}$. How do we choose $(W/L)_1$ and $(W/L)_2$ with respect to $(W/L)_{REF}$? Recall from Chapter 2 that, owing to the side diffusion of the source/drain regions, the effective channel length is less than the drawn length by $2L_D$, a poorly controlled quantity. Thus, to avoid large mismatches, the lengths of the transistors must be equal and the currents must be scaled by proper choice of the widths. We then postulate that $W_1 = 0.5W_{REF}$ and $W_2 = 2W_{REF}$. Figure 18.23 shows how M_{REF}, M_1, and M_2 in this example are laid out to ensure reasonable matching. Note that all equivalent widths are integer multiples of a unit value, W_u. Transistor M_1 is identical to M_{REF} except that half of its source remains floating (or connected to the drain). To improve the matching, the array can be surrounded by dummy devices.

18.2.4 Passive Devices

The implementation of passive devices in mainstream CMOS technologies continues to pose difficult challenges. When introduced for production, a new generation of a CMOS process provides only NMOS and PMOS devices, rarely allowing the use of polysilicon resistors (with silicide block) or high-density linear capacitors. Since it takes approximately two years to add such modules to the technology, we say "analog" processes are about two years behind "digital" processes. This is an important observation because in two years the next generation of the basic CMOS process is launched (Fig. 18.24), providing scaled transistors having a higher "raw" speed.

Which generation should a manufacturer use at $t = t_1 + 2$ years: generation NA with well-characterized passive components but a minimum dimension of L_1 or generation $N + 1$ with no high-quality passive devices but a scaled dimension of $L_1/2$? Considering the difficulties in analog design without such passive elements, we may choose generation NA, forsaking the speed advantages of generation $N + 1$. However, since the design of digital circuits is immediately moved to the scaled technology, the analog building blocks must follow suit if they are to be integrated on the same chip along with the digital system.

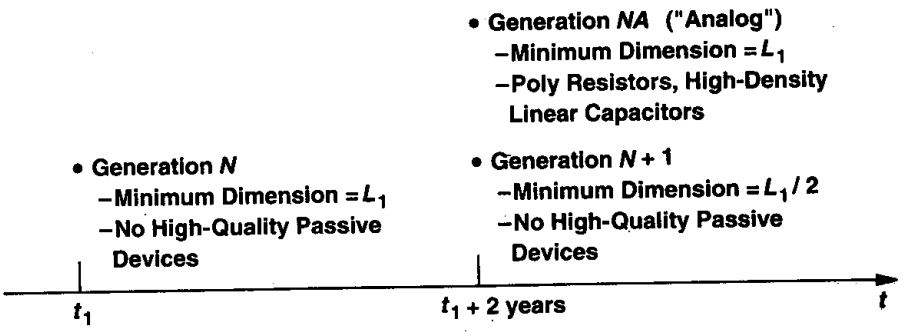


Figure 18.24 Development of digital and analog generations of a CMOS technology.

In practice, different IC manufacturers have adopted different approaches: some develop products in analog technologies, exploiting the well-characterized properties of the devices to design aggressively, whereas others utilize digital processes, taking advantage of more relaxed power-speed trade-offs and maintaining compatibility with digital circuits.

Let us now study the implementation of passive devices.

Resistors Polysilicon resistors using a silicide block exhibit high linearity, low capacitance to the substrate, and relatively small mismatches. The linearity of these resistors in fact depends on their length [1], necessitating accurate measurement and modeling for high-precision applications. Fig. 18.25 depicts an example where the nonlinearity of the resistor is critical. Since $V_{out} = -I_{in}R_F$, the accuracy of current-to-voltage conversion depends on the linearity of R_F.

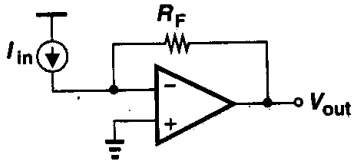


Figure 18.25 Feedback amplifier converting a voltage to current.

As with other devices, the matching of polysilicon resistors is a function of their dimensions. For example, resistors having a length of 5 μm and width of 3 μm display typical mismatches on the order of 0.2%. Most of the symmetry rules described for the layout of MOS devices apply to resistors as well. For example, resistors that are required to bear a well-defined ratio must consist of identical units placed in parallel or series (with the same orientation).

Example 18.2

Consider the bandgap circuit shown in Fig. 18.26. Choose the values of n, R_1, and R_2 such that V_{out} exhibits a zero temperature coefficient and the layout can be designed for high precision.

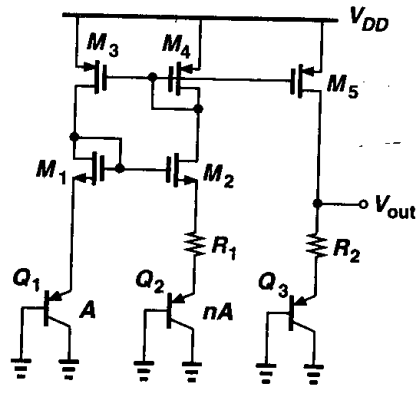


Figure 18.26

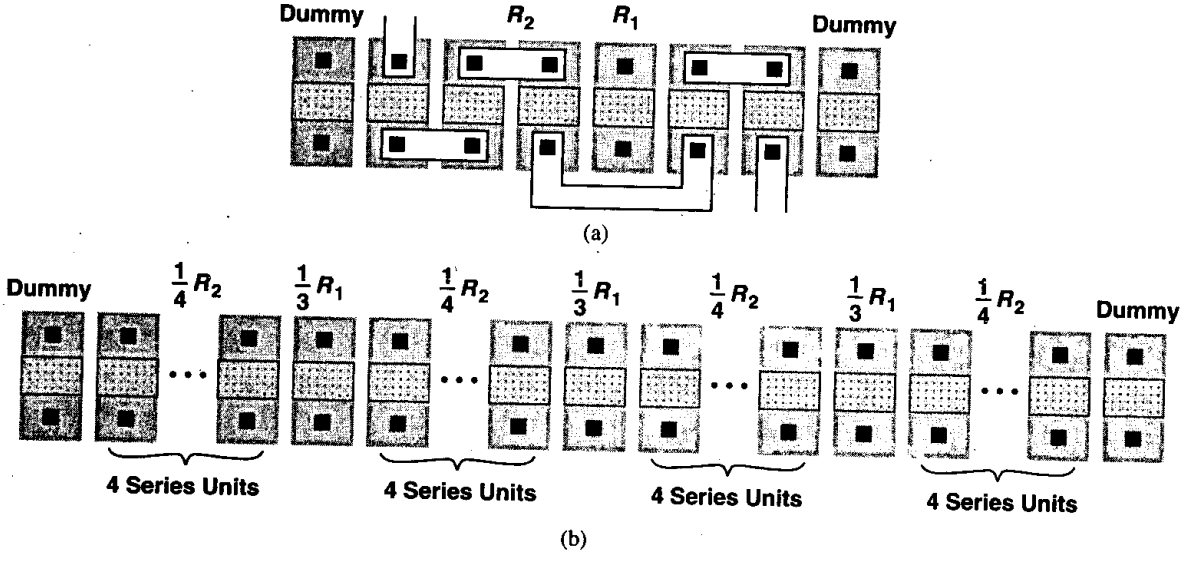


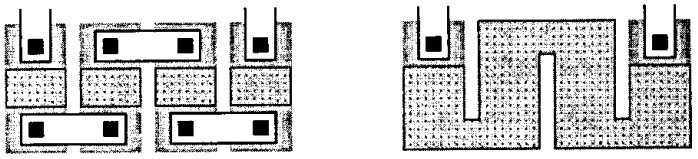
Figure 18.27 Layout of R_1 and R_2 with (a) $R_2/R_1 = 5$, (b) $R_2/R_1 = 5.34 \approx 16/3$.

Solution

Since $V_{out} = V_{BE3} + V_T(R_2/R_1)\ln n$, we must find convenient values of n, R_1, R_2 such that $(R_2/R_1)\ln n \approx 17.2$ (Chapter 11). If $n = 31$, then $R_2/R_1 \approx 5$, yielding the layout of Fig. 18.27(a). Note that R_1 is placed in the middle to partially cancel the effect of gradients.

Now suppose we choose $n = 25$, obtaining $R_2/R_1 = 5.34$. Such a value cannot be accurately established by simply adjusting the dimensions of R_2 and R_1. Rather, we write $R_2/R_1 = 16/3$ and construct the resistors as shown in Fig. 18.27(b).

For large values, resistors are usually decomposed into shorter units that are laid out in parallel and connected in series [Fig. 18.28(a)]. From the viewpoint of matching and reproducibility, this structure is preferable to "serpentine" topologies [Fig. 18.28(b)], where the corners contribute significant resistance.



(a) (b)

Figure 18.28 (a) Layout of large resistors, (b) serpentine topology.

The sheet resistance, $R_\square$, of polysilicon resistors varies with temperature and process, necessitating provisions in the design for this variation. The temperature coefficient depends on the doping type and level and must be measured for each technology. Typical values are $+0.1\%$ /°C and -0.1% /°C for p^+ and n^+ doping, respectively. The variation with process is usually less than $\pm 20\%$.

In technologies lacking a silicide block mask, resistors may be made of n-well, source/drain p^+ or n^+ material, silicided polysilicon, or metal, with $R_\square$ decreasing in this order. The sheet resistance of n-well is typically around 1 kΩ but it may vary by a large fraction, e.g., $\pm 40\%$, with process. Furthermore, $R_\square$ depends on the *width* of the resistor, as exemplified by the plot of Fig. 18.29. This is because, with a depth of several microns, n-well regions exhibit width-dependent diffusion at the edges. Also, $R_\square$ is a strong function of the n-well-substrate voltage difference, giving rise to both nonlinearity and poor

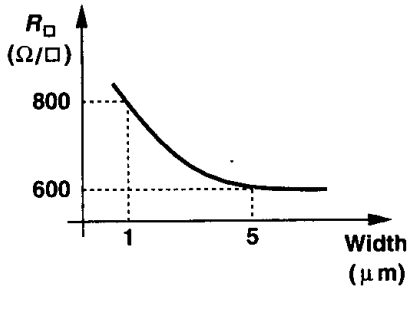


Figure 18.29 Dependence of n-well sheet resistance upon resistor width.

definition of the value of the resistor. For example, in the circuit of Fig. 18.30, resistors R_S and R_D suffer from large mismatches in $R_\square$ because the depletion region below R_S is quite narrower than that below R_D. Also, as V_{out} varies, so does the sheet resistance of R_D, introducing nonlinearity. Resistors made of n-well display a TC of $+0.2\%$ to $+0.5\%$ /°C.

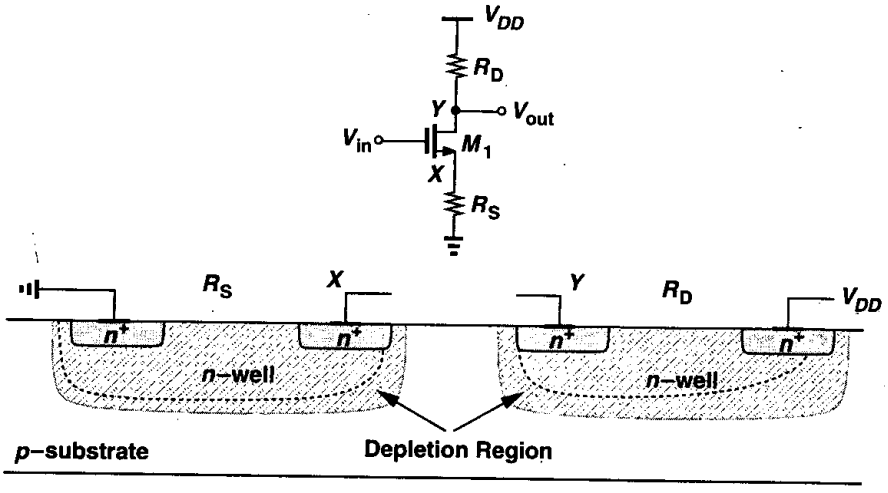


Figure 18.30 Common-source stage using n-well resistors.

Example 18.3

An A/D converter incorporates a resistor ladder consisting of 128 units made of n-well to generate equally-spaced reference voltages (Fig. 18.31). If the two ends of the ladder are connected to $V_1 = +1$ V and $V_2 = +2$ V, calculate the ratio R_{128}/R_1.

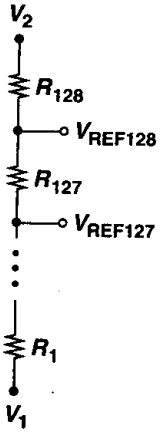


Figure 18.31 Resistor ladder used in an A/D converter.

Solution

The width of the depletion region inside the n-well is given by $x_d = \sqrt{2\epsilon_{si}(\phi_B + V_R)/(qN_{well})}$, where N_{well} denotes the n-well doping level and V_R the reverse bias voltage. Assuming the zero-bias

depth of the n-well is equal to t_0, we have

$$\frac{R_{128}}{R_1} = \frac{t_0 - \sqrt{\frac{2\epsilon_{si}}{qN_{well}}(\phi_B + V_1)} + \sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}}{t_0 - \sqrt{\frac{2\epsilon_{si}}{qN_{well}}(\phi_B + V_2)} + \sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}} \tag{18.11}$$

$$= \frac{t_0 + \sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \frac{V_1}{\phi_B}}\right)}{t_0 + \sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \frac{V_2}{\phi_B}}\right)}. \tag{18.12}$$

If the difference between R_1 and R_{128} is small, we can divide the numerator and denominator of (18.12) by t_0 and approximate the result as

$$\frac{R_{128}}{R_1} \approx \left[1 + \frac{1}{t_0}\sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \frac{V_1}{\phi_B}}\right)\right]\left[1 - \frac{1}{t_0}\sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \frac{V_2}{\phi_B}}\right)\right] \tag{18.13}$$

$$\approx 1 + \frac{1}{t_0}\sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(\sqrt{1 + \frac{V_2}{\phi_B}} - \sqrt{1 + \frac{V_1}{\phi_B}}\right). \tag{18.14}$$

For example, if $t_0 = 2\ \mu$m, $N_{well} = 10^{16}$ cm^{-1}, and $\phi_B = 0.7$ V, the mismatch between R_{128} and R_1 is nearly 60%.

The p^+ and n^+ source/drain regions can also be used as resistors. With a sheet resistance of 3 to 5 ohms per square, silicided S/D regions are suited to only low-value resistors, but their variation with process can be as high as 50%. Furthermore, the junction between these areas and the bulk introduces substantial capacitance and voltage dependence.[4]

Silicided polysilicon has a sheet resistance of 3 to 5 ohms per square and can be utilized for low resistor values. While suffering from less capacitance to the substrate than n^+ or p^+ resistors, silicided polysilicon has a process-dependent $R_\square$, with variations as high as 60 to 70%. Thus, it can be used only if its absolute value is not critical, for example, in the resistor ladder of Fig. 18.31. The temperature coefficient of this type of resistor is between +0.2 and +0.4%/°C.

The metal layers in a process can provide very low resistor values. For example, in high-speed A/D converters, the ladder of Fig. 18.31 may be constructed as simply a long metal line having equally spaced taps (Fig. 18.32). Note, however, that if the width of the metal

[4]The nonlinearity of n-well resistors is much higher because the low doping level in the n-well results in a greater sensitivity to the voltage with respect to the substrate.

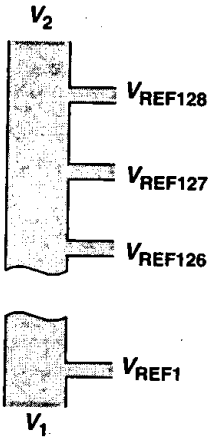


Figure 18.32 Resistor ladder made of metal.

resistor is small, matching suffers. The temperature coefficient of the resistance is about 0.3%/°C for aluminum.

Capacitors As explained in Chapter 17, high-density linear capacitors can be fabricated using polysilicon over diffusion, polysilicon over polysilicon, or metal over polysilicon, with a relatively thin layer of oxide grown between the two plates. Owing to its simplicity, the first structure is more common in today's analog processes even though it exhibits lower linearity than do the other two.

In the absence of the above structures, linear capacitors must be designed using sandwiches made of the available conductive layers. For example, in a process having four layers of metal, the capacitors can be formed as shown in Fig. 18.33. The choice of one topology over another is determined by two factors: (1) the area occupied by the capacitor and (2) the ratio of the bottom-plate parasitic capacitance to the interplate capacitance, C_P/C. In typical technologies, the capacitance between consecutive metal layers [e.g., $C_1, \ldots, C_4$ in Fig. 18.33(d)] is on the order of 35 to 40 aF/μm^2 and that between metal 1 and polysilicon is about 60 aF/μm^2. Thus, the structure of Fig. 18.33(d) provides more than four times the density of that in Fig. 18.33(a). On the other hand, the value of C_P increases from Fig. 18.33(a) to Fig. 18.33(d). With typical values, C_P/C reaches a minimum—about 0.2 to 0.25—for the structure of Fig. 18.33(a) or (b) and increases to about 0.5 for the sandwich of Fig. 18.33(d).

Since the absolute value of interlayer capacitances is poorly controlled in digital technologies, the capacitors of Fig. 18.33 may experience process variations as high as 20%. By contrast, the gate oxide capacitance is typically controlled with less than 5% error. Interestingly, the structure of Fig. 18.33(d) may suffer from less variation than the others because random variations in the capacitances between various layers tend to "average out."

We have thus far neglected the fringe capacitance. As depicted in Fig. 18.34, the electric field lines emanating from the edge of each plate must terminate on the edge of the other plate or on the substrate, giving rise to a fringe capacitance that must be taken into account. The fringe capacitance can be calculated using Eq. (17.11) or from tabulated values in the process design manual.

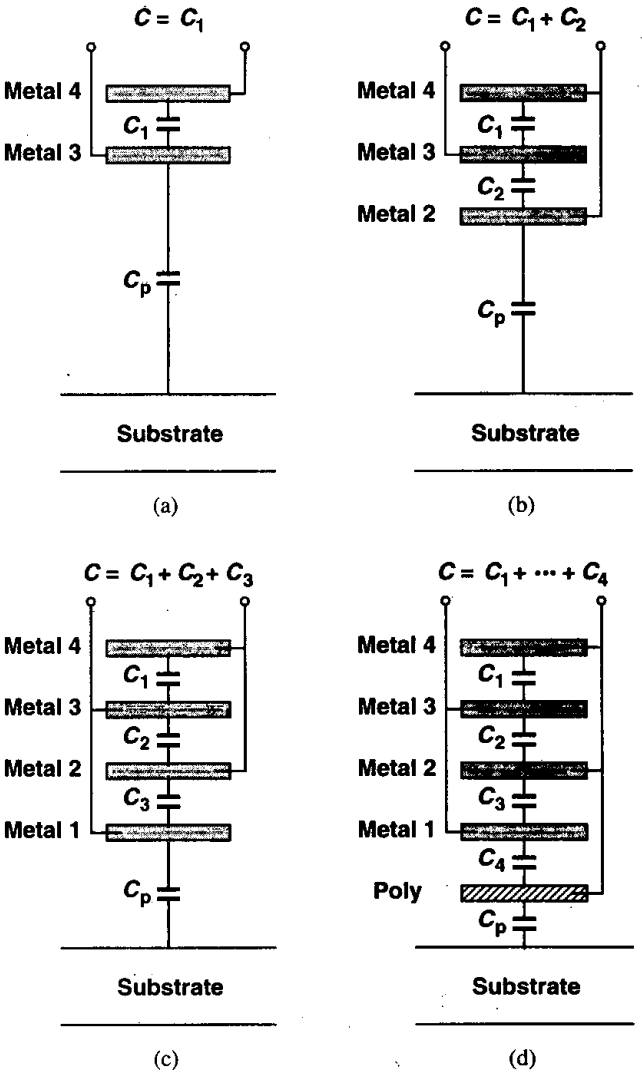


Figure 18.33 Capacitor structures using various conductive layers.

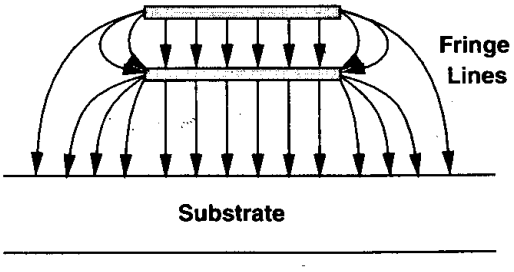


Figure 18.34 Fringe component of capacitance.

As explained in Chapter 17, a MOS transistor with its source and drain tied together can act as a capacitor if the gate-source potential is sufficient to establish an inversion layer. However, the voltage dependence of the capacitance limits the use of this structure.

The layout of capacitors for high-precision circuits must follow the principles described above for transistors and resistors. For example, in applications where an array of well-matched capacitors is required, dummy devices must be placed on the perimeter of the array.

Example 18.4

The circuit of Fig. 18.35(a) is designed for a nominal gain of $C_1/C_2 = 8$. How should C_1 and C_2 be laid out to ensure precise definition of the gain?

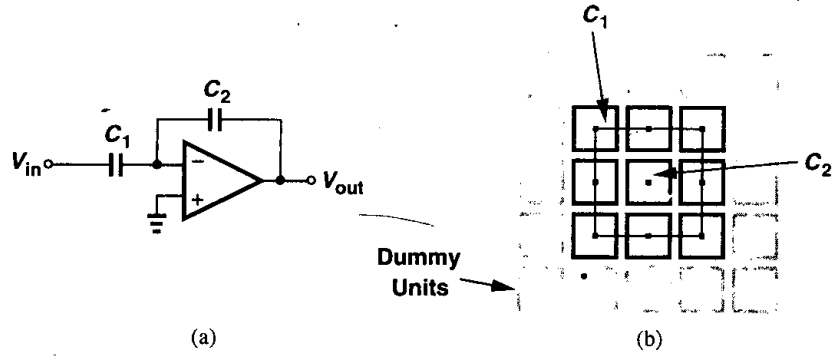


Figure 18.35

Solution

We form C_1 as 8 unit capacitors, each equal to C_2, and place all of the units in a square array [Fig. 18.35(b)]. Note that (1) C_2 is symmetrically surrounded by the units comprising C_1 so that the effect of vertical or horizontal gradients is cancelled to the first order; (2) dummy capacitor units are placed around the main array, creating approximately the same environment for the units of C_1 as that seen by C_2.

For large capacitor arrays, cross-coupling techniques such as those illustrated in Figs. 18.20 and 18.26 can be applied. However, unlike transistors and resistors, capacitors are quite sensitive to the wiring capacitance, demanding great care in the interconnection of the units. Even in the simple array of Fig. 18.35(b), it is difficult to route all of the top-plate and bottom-plate connections while introducing no additional capacitance. As the layout of Fig. 18.36 exemplifies, the wiring inevitably leads to some error in the ratio C_1/C_2.

Diodes Two types of *pn* junctions can be formed in a standard CMOS technology: one in the *p*-substrate and another in an *n*-well (Fig. 18.37). The former must remain reverse biased and can therefore serve only as a voltage-dependent capacitor ("varactor"), e.g., in voltage-controlled oscillators.

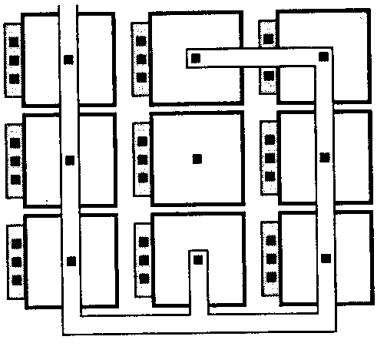


Figure 18.36 Layout of capacitors along with interconnections.

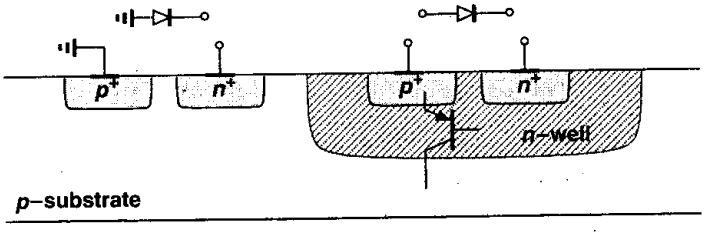


Figure 18.37 Diodes in CMOS technology.

The diode formed in an n-well also faces difficulties if forward biased. Recall from Chapter 11 that the p^+ region in the n-well, the n-well itself, and the p-substrate constitute a bipolar pnp transistor whose collector is typically grounded. Thus, if the pn junction in the n-well is forward biased, substantial current flows from the p^+ terminal to the substrate. In other words, the structure must not be viewed as merely a two-terminal floating diode. Nonetheless, if reverse-biased, the device can serve as a varactor.

Owing to these difficulties, analog CMOS circuits rarely incorporate forward-biased diodes.

18.2.5 Interconnects

Modern CMOS processes offer five metal layers for interconnection. By comparison, as late as 15 years ago, CMOS technologies provided only one layer of metal. Nevertheless, many effects related to wires must still be taken into account when a high-precision and/or high-speed circuit is laid out.

The parallel-plate and fringe capacitance of wires may degrade the speed if long interconnects are required. For example, in a mixed-signal system (e.g., using many switched-capacitor circuits), the clock signal must be distributed over long wires to access various building blocks, thereby experiencing significant line capacitance. More importantly, the capacitance between lines introduces substantial coupling of signals.

Fig. 18.38 illustrates an example of cross-talk between signals. Here, a common-source stage and a NAND gate are located next to each other and the two inputs to the gate, V_A and V_B, cross over the analog signal, V_{in}. Furthermore, the clock wire, CK, is laid out in parallel with V_{in} and the output of the NAND gate has some overlap with the output

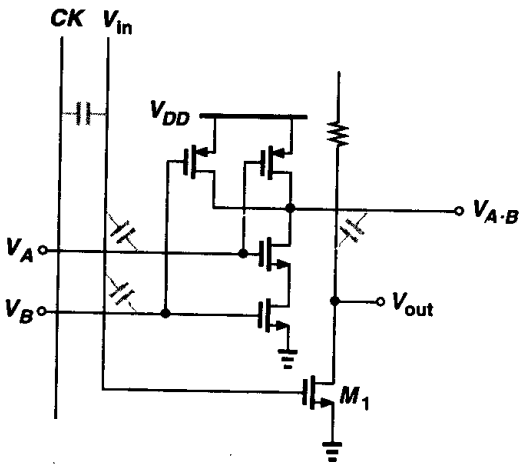


Figure 18.38 Capacitive coupling between various lines in a typical layout.

of the common-source stage. Each of the coupling capacitances in this layout may considerably corrupt V_{in} or V_{out}. Note that, even though the coupling capacitances are small, the signal corruption may be appreciable because typical voltage swings on V_A, V_B, $V_{A\cdot B}$, and CK are quite large. For example, if the overlap of V_A and V_{in} gives a capacitance of 50 aF, and the total capacitance seen from V_{in} to ground is 50 fF, then a 3-V change in V_A may result in a 3-mV corruption at V_{in}.

Crosstalk can be reduced through the use of two techniques. First, differential signaling converts most of the crosstalk to common-mode disturbance. For example, if the circuit of Fig. 18.38 is modified to that shown in Fig. 18.39, the coupling of V_A and V_B to V_{in}^+ and V_{in}^-

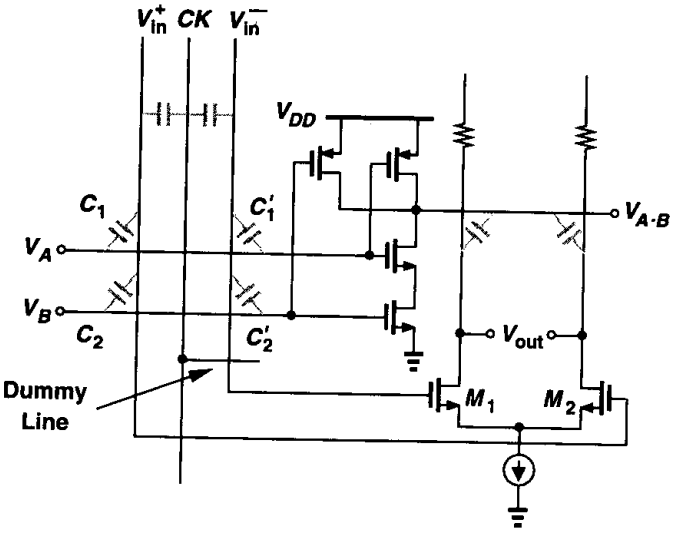


Figure 18.39 Reduction of capacitive coupling through the use of differential signaling.

produces no differential error if $C_1 = C_1'$ and $C_2 = C_2'$. Even for 10% mismatch between the capacitances, the differential corruption is one order of magnitude less than that in Fig. 18.38. Note that a dummy wire is added to the layout so as to create an overlap capacitance between CK and V_{in}^- equal to that between CK and V_{in}^+. As mentioned in Chapter 4, it is desirable to employ differential clocks as well to suppress the net coupling further.

Second, sensitive signals can be "shielded" in the layout. Depicted in Fig. 18.40(a), one approach places ground lines on the two sides of the signal, forcing most of the electric field lines emanating from the "noisy" lines to terminate on ground rather than on the signal. Note that this method proves more effective than simply allowing more space between the signal and the noisy lines [Fig. 18.40(b)]. The shielding, however, is obtained at the cost of more complex wiring and greater capacitance between the signals and ground.

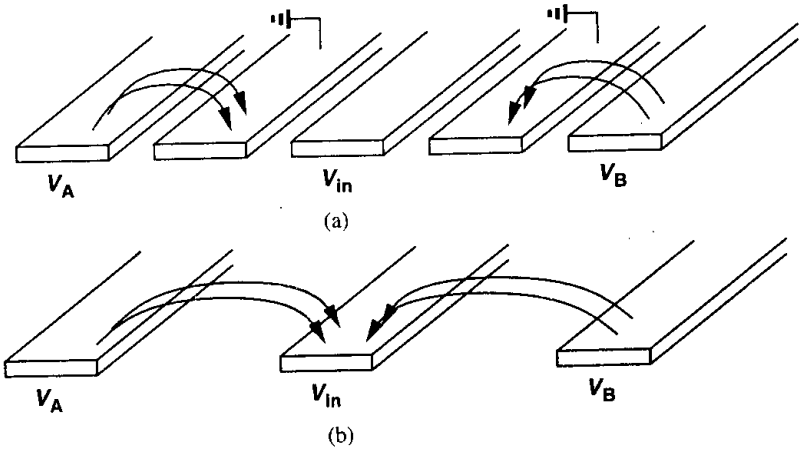


Figure 18.40 (a) Shielding sensitive signals by additional ground lines, (b) greater spacing between lines to reduce coupling.

Another shielding technique is shown in Fig. 18.41. Here, the sensitive line is surrounded by a grounded shield consisting of a higher and a lower metal layer and hence fully isolated from external electric field lines.[5] However, the signal experiences higher capacitance to ground and the use of three metal layers here complicates the routing of other signals.

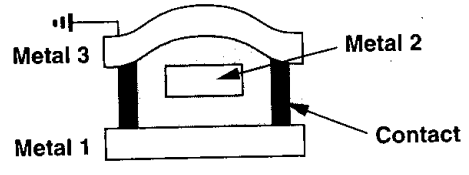


Figure 18.41 Shielding a sensitive line (metal 2) by lower and upper ground planes.

The resistance of interconnects also requires attention. In low-noise applications, long signal wires—with sheet resistances of 40 to 80 m$\Omega/\square$—may introduce substantial thermal noise. Furthermore, the contacts and vias also suffer from a high resistance. For example,

[5] We assume that the ground connection itself does not contain noise. We return to this issue in Section 18.4.

a 0.3-μm $\times$ 0.3-μm metal contact to silicided polysilicon exhibits a resistance of 5 to 10 Ω and a via between metal 1 and metal 2, a resistance of 5 Ω.

Example 18.5

In the layout of Fig. 18.42, a 100-μm metal 4 line is connected to a sequence of vias and contacts to reach the gate of a transistor. Calculate the thermal noise contributed by the line and the contacts.

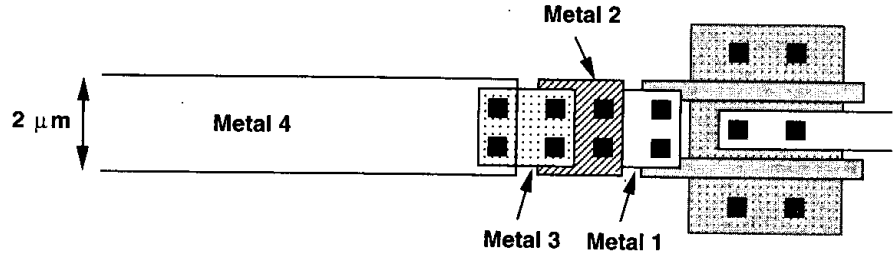


Figure 18.42

Solution

Assuming $R_\square = 40\ \text{m}\Omega/\square$ for metal 4, a via resistance of 5 Ω, and a contact resistance of 10 Ω, we have $R_{tot} = 2+2.5+2.5+2.5+5 = 14.5\ \Omega$. The thermal noise voltage is thus equal to $0.49\ \text{nV}/\sqrt{\text{Hz}}$ at room temperature.

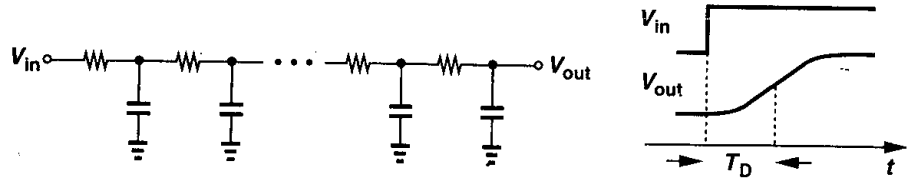


Figure 18.43 Delay and dispersion of a signal in a long line.

The distributed resistance and capacitance of long interconnects may introduce significant delay and "dispersion" in signals. Illustrated in Fig. 18.43, the delay can be approximated as

$$T_D = \frac{1}{2} R_u C_u L^2, \tag{18.15}$$

where R_u and C_u denote the resistance and capacitance per unit length, respectively, and L is the total length. For example, consider the circuit shown in Fig. 18.44, where an array of samplers senses the analog input V_{in} and is activated by CK. If the delays experienced by CK and V_{in} from the left side to the right side are not equal, then the levels sampled by $C_1, \ldots, C_n$ are not equal, resulting in distortion in the sampled waveform. Even if the clock

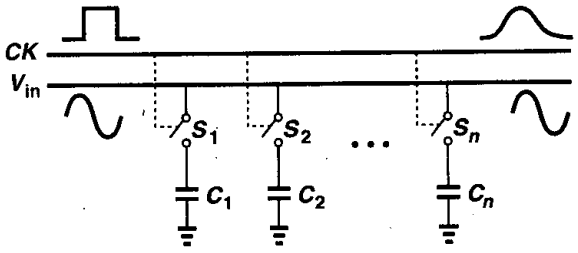


Figure 18.44 An array of sampling circuits sensing an input.

and signal lines and their capacitive loading are identical, CK and V_{in} may still suffer from unequal delays because the former is a rectangular wave and the latter is not.

The term "dispersion" refers to the significant increase in the transition time of the signal as it propagates through a line, a particularly troublesome effect if a clock edge is to define a sampling point. In the example of Fig. 18.44, the clock waveform applied to S_n displays long rise and fall times, making the sampling susceptible to both noise and distortion [4]. The clock edges can be sharpened by inserting an inverter between CK and every switch but at the cost of greater uncertainty in the delay difference between CK and V_{in}.

As mentioned in Chapter 17, the design of power and ground busses on a chip requires attention to a number of issues. In large ICs, the dc or transient voltage drop along the busses may be significant, affecting sensitive circuits supplied by the same lines. Furthermore, electromigration mandates a minimum line width to guarantee long-term reliability. With multiple interconnect levels available in today's CMOS technology, it is possible to connect two or more layers in parallel, thereby reducing the series resistance and alleviating electromigration constraints. Since the thickness of the top metal layer is typically twice that of the lower ones, at least *three* layers must be placed in parallel to relax these issues by a factor of two. As a result, routing signals and bias lines across the busses may become difficult if only one or two more layers of metal are available.

If the bias currents drawn from a long bus are relatively well-defined, then the bus width can be "tapered" from one end to the other so as to create a relatively constant voltage drop along the line. Illustrated in Fig. 18.45, this technique can be used if the metal resistance and its temperature coefficient are known.

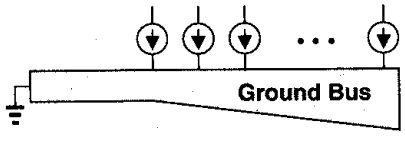


Figure 18.45 Tapered ground line for reduction of voltage drops.

18.2.6 Pads and ESD Protection

The interface between an integrated circuit and the external environment involves a number of important issues. In order to attach bond wires to the die, large "pads" are placed

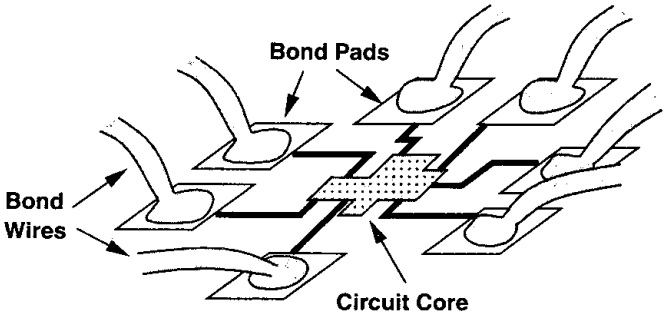


Figure 18.46 Addition of bonding pads to a chip.

on the perimeter of the chip and connected to the corresponding nodes in the circuit (Fig. 18.46).

The pad dimensions and structure are dictated by the reliability issues and margin for manufacturing tolerances in the wire bonding process. With bond wire diameters ranging from 25 μm to 50 μm, the minimum pad size falls between roughly 70 μm $\times$ 70 μm and 100 μm $\times$ 100 μm. Adjacent pads are usually separated by at least 25 μm. From the circuit design point of view, the pad dimensions must be minimized so as to reduce both the capacitance of the pad to the substrate and the total die area.

A simple pad would consist of only a square made of the top metal layer. However, such a structure is susceptible to "lift-off" during bonding. For this reason, each pad is typically formed by the two topmost metal layers, connected to each other by many small vias on the perimeter (Fig. 18.47). Note that this structure suffers from a larger capacitance to the substrate than a pad made of only the top layer.

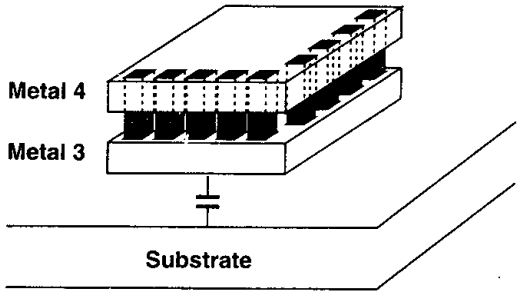


Figure 18.47 Structure of a typical bonding pad.

Example 18.6

Calculate the capacitance of a metal-4 pad and a metal-4/metal-3 pad. Assume dimensions of 75 μm $\times$ 75 μm and use the capacitance data shown in Fig. 18.48.

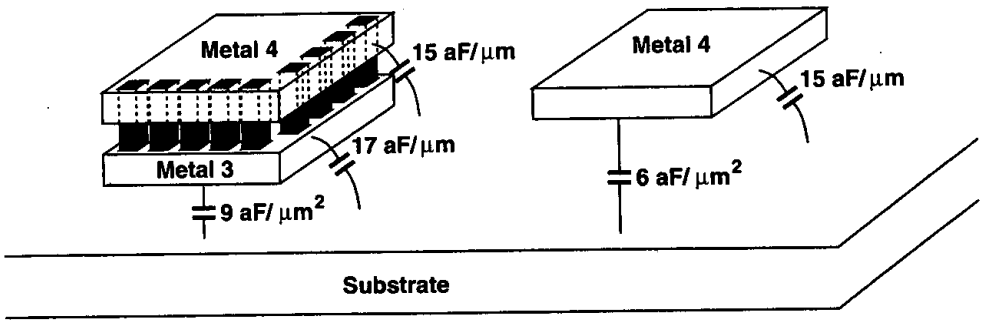


Figure 18.48

Solution

For a metal-4 pad,

$$C_{tot} = 75^2 \times 6 + 75 \times 4 \times 15 \tag{18.16}$$

$$= 38.25 \text{ fF}. \tag{18.17}$$

For a metal-4/metal-3 pad,

$$C_{tot} = 75^2 \times 9 + 75 \times 4 \times (17 + 15) \tag{18.18}$$

$$= 60.22 \text{ fF}. \tag{18.19}$$

Note that the fringe capacitances of metal 4 and metal 3 are directly added here. This is a rough approximation.

The interface between an IC and the external world also entails the problem of electrostatic discharge (ESD). This effect occurs when an external object having a high potential touches one of the connections to the circuit. Since the capacitance seen at each input or output is quite small, the ESD produces a large voltage, possibly damaging the devices fabricated on the chip.

A common case of ESD arises when ICs are handled by human beings. For this effect, the human body can be modeled by a capacitance of a few hundred picofarads in series with a resistance of a few kiloohms. Depending on the environment, the voltage across the capacitance ranges from a few hundred volts to several thousand volts. Thus, if a person touches a line connecting to the chip, the chip is easily damaged. Interestingly, electrostatic discharge may occur even without actual contact because at high electric fields, the person's finger "arcs" to the connection through the air if the finger is sufficiently close to the line.

It is important to note that ESD may occur even without human intervention. If not properly grounded, various objects in a typical chip assembly line accumulate charge, rising to high potential levels. Furthermore, charge in dry air may create substantial potential gradients with respect to ground.

MOS devices sustain two types of permanent damage as a result of ESD. First, the gate oxide may break down if the electric field exceeds roughly 10^7 V/cm (e.g., 10 V for an

oxide thickness of 100 Å), typically leading to a very low resistance between the gate and the channel. Second, the source/drain junction diodes may melt if they carry a large current in forward or reverse bias, creating a short to the bulk. For today's short-channel devices, both of these phenomena are likely to occur.

In order to alleviate the problem of electrostatic discharge, CMOS circuits incorporate ESD protection devices. Illustrated in Fig. 18.49, such devices clamp the external discharge to ground or V_{DD}, thereby limiting the potential applied to the circuit. Resistor R_1 is usually necessary so as to avoid damaging D_1 or D_2 due to large currents that would otherwise flow from the external source.

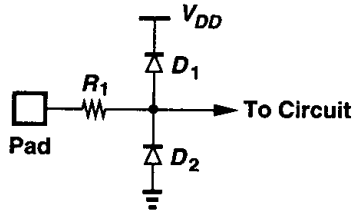


Figure 18.49 Simple ESD protection circuit.

The use of ESD protection structures involves three critical issues. First, the devices introduce substantial capacitances from the node to ground and V_{DD}, degrading the speed and the matching of impedances at the input and output ports of the circuit. Since the protection devices, e.g., D_1 and D_2 in Fig. 18.49, must be large enough so that the chip sustains a high ESD voltage without damage, their capacitance may reach several picofarads. The thermal noise of R_1 may also become significant.

Second, the parasitic capacitance of the ESD devices may couple noise on V_{DD} to the input of the circuit, corrupting the signal. We return to this issue in Section 18.4.

Third, if not properly designed, ESD structures may lead to latchup in CMOS circuits when electrostatic discharge occurs during actual circuit operation (or even when the circuit is turned on). For this reason, process engineers fabricate and characterize many different ESD structures for each generation of a technology, eventually providing a few reliable configurations that can be used in circuits.[6]

18.3 Substrate Coupling

Most modern CMOS technologies use a heavily-doped p^+ substrate to minimize latchup susceptibility. However, the low resistivity of the substrate (on the order of 0.1 $\Omega \cdot$ cm) creates unwanted paths between various devices in the circuit, thereby corrupting sensitive signals. Called "substrate coupling" or "substrate noise," this effect has become a serious issue in today's mixed-signal ICs [2].

To understand this phenomenon, suppose a CMOS inverter sensing a clock is laid out next to a common-source stage amplifying an analog signal [Fig. 18.50(a)]. Note that the

[6]In general, a circuit designer should not use an ESD structure that has not been tested and qualified for the technology. Uncharacterized ESD devices are likely to cause latchup.

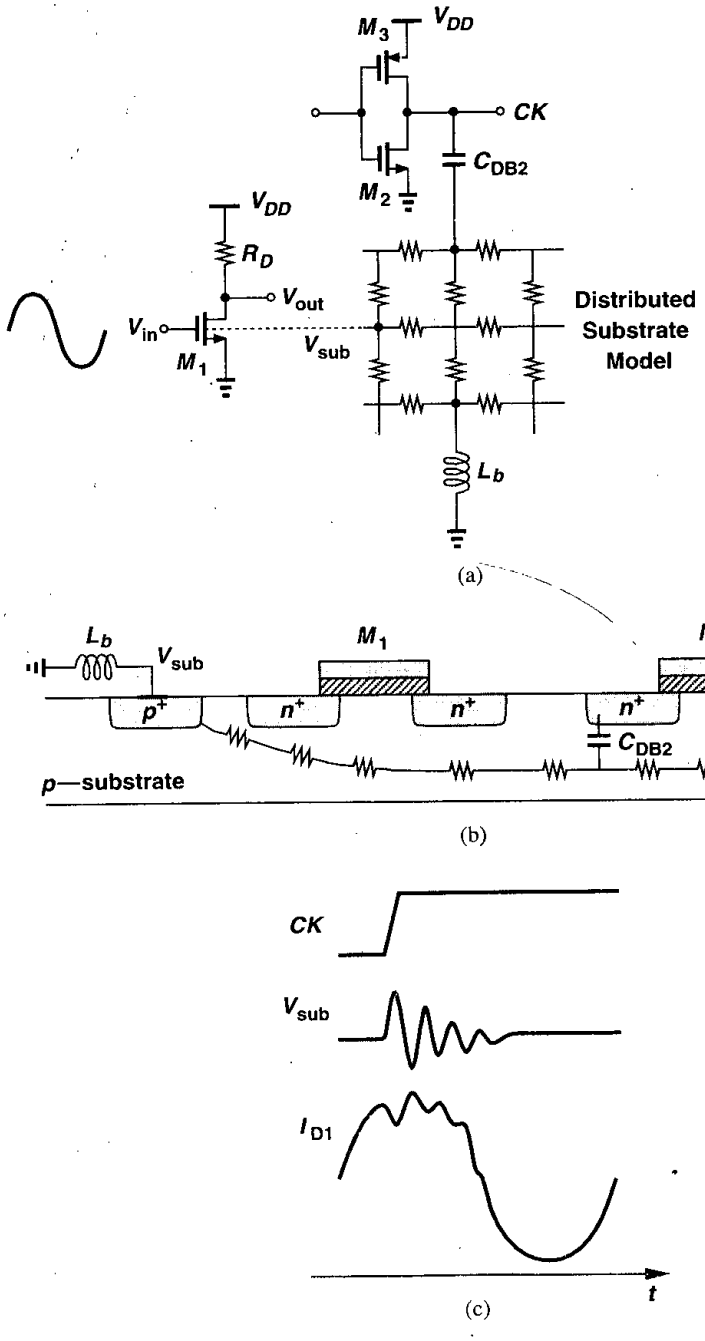


Figure 18.50 (a) Mixed-signal circuit including the effect of substrate coupling, (b) sideview of device layout, (c) signal waveforms.

substrate is connected to ground through a bond wire that exhibits an (unwanted) inductance of L_b. With the aid of the cross section depicted in Fig. 18.50(b), we observe that the large voltage excursions at the drain of M_2 are coupled to the substrate through the drain junction capacitance, disturbing the substrate voltage because of the finite impedance of L_b.

How does the substrate noise influence M_1? The principal coupling mechanism here occurs through body effect, varying the threshold voltage of M_1 with the substrate voltage. Since the drain current of M_1 depends on $V_{in} - V_{TH1}$, variations in V_{TH1} are indistinguishable from those in V_{in}. In other words, as illustrated in Fig. 18.50(c), every transition of CK disturbs the analog output.

The problem of substrate coupling becomes more noticeable as the number of "noise" generators increases. In a mixed-signal environment, thousands of digital gates may inject noise into the substrate—especially during clock transitions—introducing hundreds of millivolts of disturbance in the substrate potential. The disturbance is also proportional to the size of the noise-injecting devices, an important issue if large transistors are used as buffers driving heavy external loads.

It may seem that substrate coupling can be decreased by increasing the physical spacing between sensitive building blocks and digital sections of a chip. In practice, however, this remedy may not be effective or feasible. If heavily doped, the substrate operates as a low-resistance plane, distributing a relatively uniform potential across the chip regardless of the position of the noise generators [3]. Furthermore, in many mixed-signal systems, the analog and digital functions are so heavily blended that it is difficult to separate their corresponding circuits. Fig. 18.51 shows a slice of an A/D converter consisting of a comparator, a flipflop, a NAND gate, and a read-only memory (ROM). Various logical swings in the comparator and the digital circuits generate substrate noise, but increasing the distance between any two blocks necessitates long interconnects, degrading the performance.

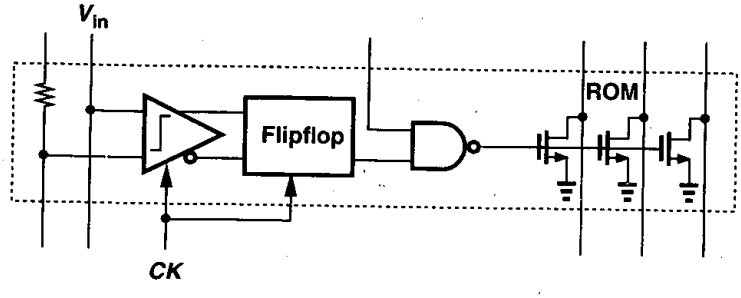


Figure 18.51 A slice of an A/D converter.

In order to minimize the effect of substrate noise, the following methods can be applied. First, differential operation should be used throughout the circuit, making the analog section less sensitive to common-mode noise. Second, digital signals and clocks should be distributed in complementary form, thereby reducing the net amount of the coupled noise. Third, critical operations, e.g., sampling a signal or transferring charge from one capacitance to another, should be performed well after clock transitions such that the substrate voltage settles. Fourth, the inductance of the bond wire connected to the substrate should

be minimized (Section 18.4). Also, op amps using a PMOS differential input are preferred because the well of the transistors can be tied to their common source, reducing the effect of substrate noise.

In circuits fabricated on lightly-doped substrates, "guard rings" can be employed to isolate the sensitive sections from the substrate noise produced by other sections. A guard ring may be simply a continuous ring made of substrate ties that surrounds the circuit, providing a low-impedance path to ground for the charge carriers produced in the substrate. With its large depth, the n-well can also augment the operation of a guard ring by stopping the noise currents flowing near the surface (Fig. 18.52).

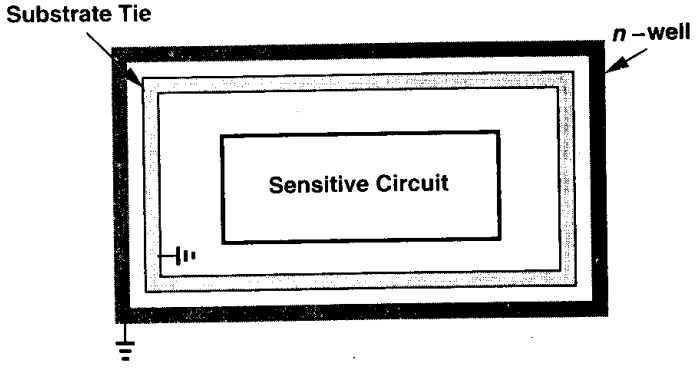


Figure 18.52 Use of guard ring to protect sensitive circuits.

In large mixed-signal ICs, it may not be possible to avoid substrate "bounce" with respect to the external ground because of the high transient currents drawn by the devices and the finite impedance of the bond wire connected to the substrate. However, we recognize that if the ground of the chip bounces in unison with the substrate, then the transistors experience no noise. Illustrated in Fig. 18.53, this idea suggests that the ground and the substrate should be connected on the chip and brought out through a single wire.

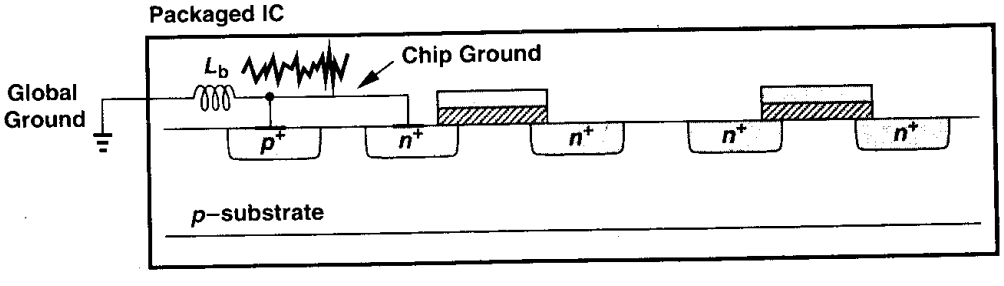


Figure 18.53 Substrate bounce.

The connection of the substrate to the chip ground nonetheless faces two difficulties. The first relates to "ground bounce." As shown in Fig. 18.54 and explained in Section 18.4,

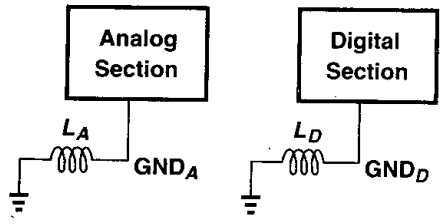


Figure 18.54 Analog and digital grounds.

most mixed-signal circuits employ at least one "analog ground" and one "digital ground" so as to avoid corrupting the analog section by the large transient noise produced by the digital section. To which ground should the substrate be connected? If the analog ground is used, then the large substrate noise current must flow through L_A, creating noise on GND_A [Fig. 18.55(a)], and if the digital ground is used then the substrate voltage is heavily disturbed by the large noise on GND_D [Fig. 18.55(b)]. Of course, connecting the substrate to both GND_A and GND_D gives rise to a low-resistance path between the two, defeating the purpose of separating the analog and digital grounds.

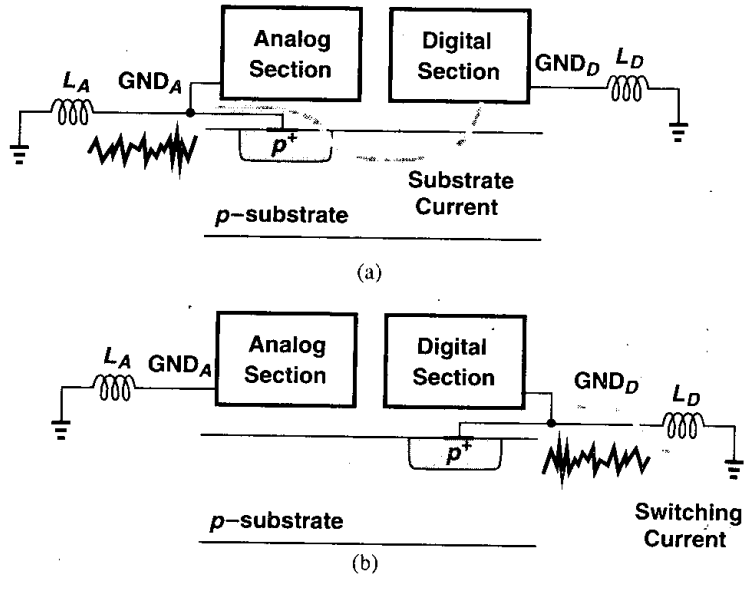


Figure 18.55 Connection of substrate contact to (a) analog ground, (b) digital ground.

The choice between the configurations shown in Figs. 18.55(a) and (b) depends on the transient currents drawn by the digital section from the substrate and the ground as well as the magnitudes of L_A and L_D. In most cases, the topology of Fig. 18.55(a) is preferred because it ensures the analog ground voltage and the substrate potential vary in unison. As illustrated in Fig. 18.56(a), if the analog ground and the substrate experience unequal bounce, then the drain current of M_1 is corrupted by the substrate noise. The configuration

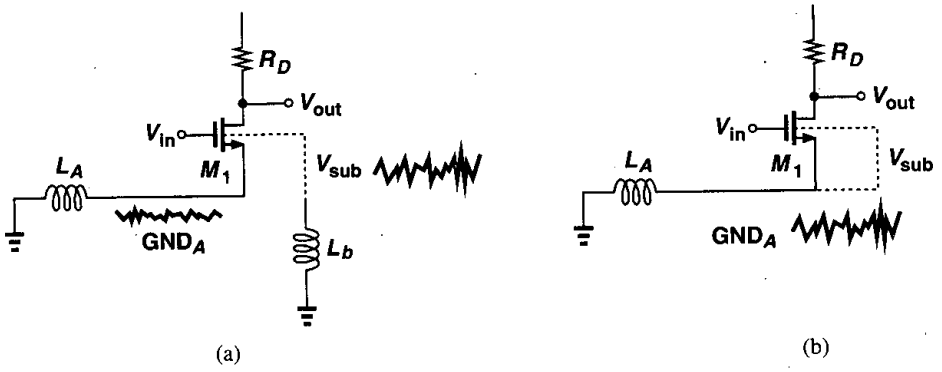


Figure 18.56 (a) Large source-bulk noise voltage due to separating substrate contact from analog ground, (b) suppression of the effect.

of Fig. 18.56(b), on the other hand, introduces less noise in I_{D1}. In general, careful, realistic simulations of the overall environment (including the package) are necessary to determine which approach yields less noise.

The second issue in allowing the substrate and a chip ground to bounce together is the difficulty in defining a reference potential for the input signals. As shown in Fig. 18.57(a), a single-ended input is heavily corrupted as its reference point changes from the off-chip ground to the on-chip ground. For the differential structure of Fig. 18.57(b), the effect is much less pronounced but in high-precision applications, asymmetries in the circuit and interconnections convert a fraction of the common-mode noise to a differential component.

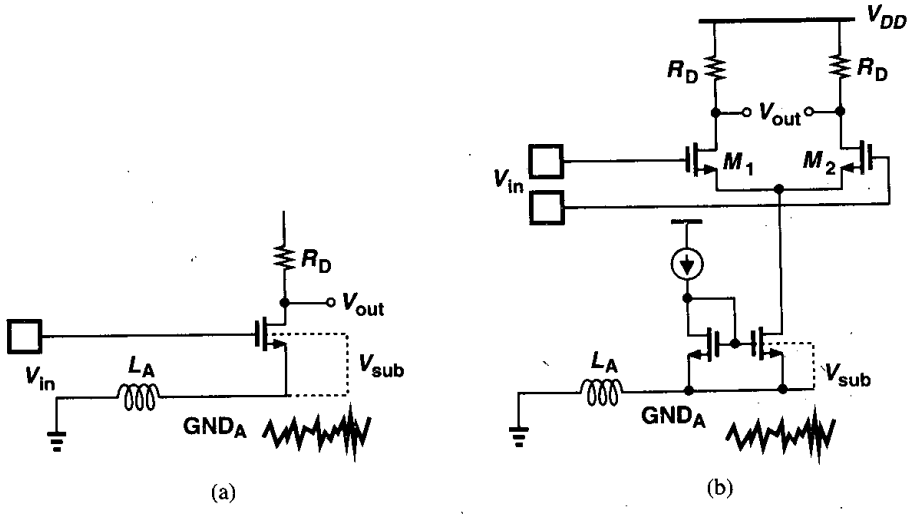
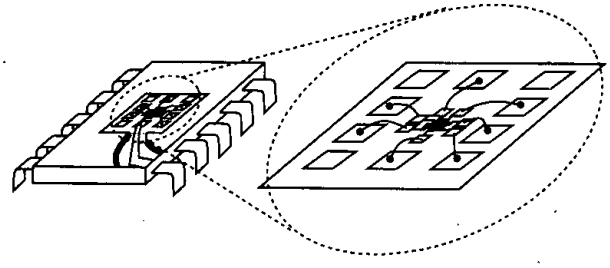


Figure 18.57 (a) Input signal corruption due to ground and substrate bounce, (b) less corruption in a differential environment.

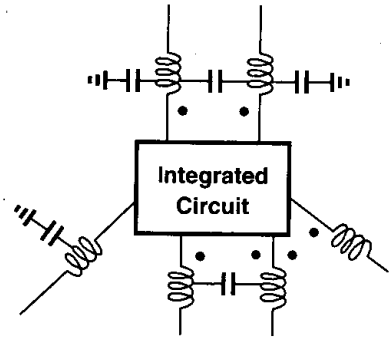
18.4 Packaging

After fabrication and dicing, integrated circuits are packaged. The parasitics associated with the package and connections to the chip introduce many difficulties in the evaluation of the actual performance of the circuit at high speeds and/or high accuracies.

Let us first consider a simple dual-in-line package (DIP) [Fig. 18.58(a)]. Here, the die is mounted in the center cavity and bonded to the pads on the perimeter of the cavity. These pads are in fact the tip of each trace that ends in each package pin. Such a structure exhibits the following parasitics: bond wire self-inductance, trace self-inductance, trace-to-ground capacitance, trace-to-trace mutual inductance, and trace-to-trace capacitance. Thus, as shown in Fig. 18.58(b), the connections between the circuit and the external world are far from ideal.



(a)



(b)

Figure 18.58 (a) Dual-in-line package, (b) electrical model of the package.

While, owing to both circuit innovations and device scaling, the speed and accuracy of integrated circuits have steadily increased, the performance of packages, especially for low-cost applications, has not improved significantly. This limitation originates from the unscalable nature of packages and the environment in which they are used. For example,

the diameter of the bond wires, the width and spacing of package pins, and the width and spacing of the traces in printed circuit (PC) boards are determined by mechanical stress, ease and cost of assembly, series resistance at high frequencies (skin effect), etc. In the past 20 years, these dimensions have scaled by less than a factor of five whereas the speed of many mixed-signal circuits has increased by two orders of magnitude. As a result, packaging continues to limit the achievable performance of today's high-performance ICs.

The foregoing difficulties mandate that the package parasitics be taken into account in the design of integrated circuits—sometimes from the very beginning. Thus, simulations must include a reasonable circuit model of the package, and the design and layout must take many precautions to minimize the effect of package parasitics.

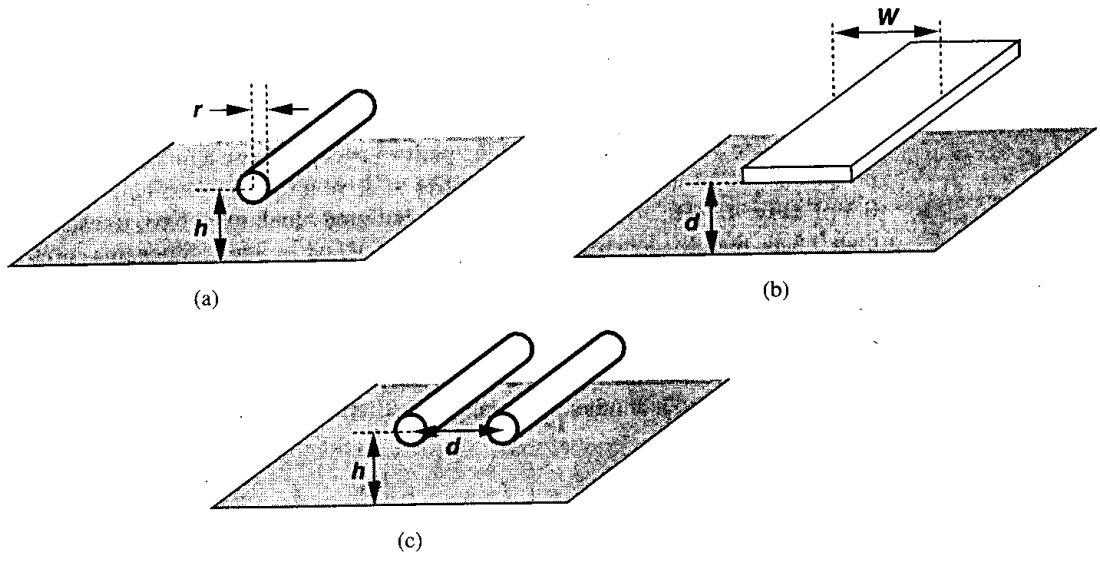


Figure 18.59 Common geometries in packaging.

Since many package manufacturers do not provide circuit models for their products, IC designers often develop the models themselves by calculations and measurements. Fig. 18.59 depicts three common cases of self- and mutual inductance. From [6], we have for a round wire above a ground plane [Fig. 18.59(a)]:

$$L \approx 0.2 \ln \frac{2h}{r} \text{ nH/mm}, \tag{18.20}$$

which amounts to roughly 1 nH/mm for typical bond wires. For a flat trace above a ground plane [Fig. 18.59(b)]:

$$L \approx \frac{1.6}{K_f} \cdot \frac{d}{W} \text{ nH/mm}, \tag{18.21}$$

where K_f denotes the fringe factor and from the data in [6] can be approximated as

$0.72(d/W) + 1$. For two round wires above a ground plane, the mutual inductance is [6]

$$L_m = 0.1 \ln\left[1 + \left(\frac{2h}{d}\right)^2\right] \text{ nH/mm}. \tag{18.22}$$

The parasitic capacitances can be calculated with the simple interplate equation and Eq. (17.11).

Let us now study the effect of each type of package parasitic. We categorize the connections to the chip into five groups: power and ground lines, analog and clock inputs, outputs, reference lines, and substrate connection(s).

Self-Inductance Each bond wire and its corresponding package trace exhibit a finite self-inductance, with a total value between approximately 2 nH and 20 nH depending on the length of the wire and the type of the package. To understand how the self-inductance of supply and ground lines impacts the performance, suppose a mixed-signal circuit incorporates a CMOS inverter as a clock buffer to drive a moderate on-chip capacitance, e.g., 0.5 pF (Fig. 18.60). Also, assume that the buffered clock must have transition times less than 0.5 ns, thereby demanding a current of $C\Delta V/\Delta t = 3$ mA. Since this current is drawn from V_{DD1} and GND_1 in 0.5 ns, we can estimate the voltage drop across L_D or L_G as[7] $L\Delta I/\Delta t = 6 \times 10^6 L$. For example, if $L_D = L_G = 5$ nH, then the transient voltage across each inductor equals 30 mV. This effect is called supply and ground "bounce" or "noise." Note that if the inverter is replaced by a differential pair, the supply bounce decreases substantially (why?), another advantange of differential operation.

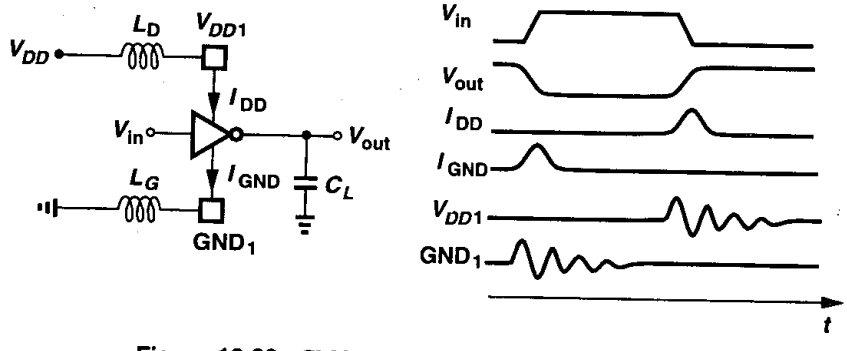


Figure 18.60 CMOS inverter driving a load capacitance.

A supply noise of 30 mV may seem quite benign, especially if the analog circuits feeding from the same supply line are fully differential. However, in a typical mixed-signal IC, hundreds or thousands of digital gates may switch during each clock transition, creating enormous noise on their supply and ground connections. For this reason, most such

[7]This calculation is quite rough because the current produced by the buffer varies during the transition.

systems employ separate supply and ground lines for the analog and digital sections, hence the terminology "analog supply" and "digital supply."

Separating power lines into analog and digital groups is not always straightforward. As an example, suppose a sampling circuit is clocked by an inverter (Fig. 18.61). Should the inverter be supplied from analog or digital power lines? If the inverter is connected to the digital supply, then the large noise on V_{DD} couples through the gate-drain overlap capacitance of M_1, corrupting V_{out} when the transistor is off. On the other hand, if many such inverters are supplied from the analog V_{DD}, they collectively draw large transient currents, corrupting the supply voltage. These cases may require a third type of power line so that it remains less noisy than the digital supplies.

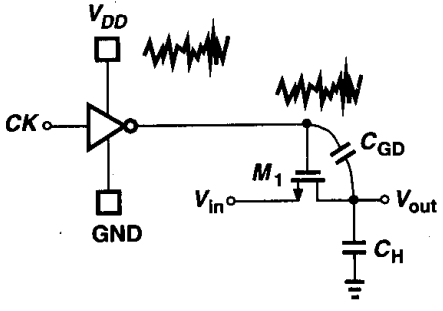


Figure 18.61 Noise in a sampling circuit resulting from the clock buffer's supply bounce.

For characterization and troubleshooting purposes, it is sometimes desirable to monitor the supply noise. Figure 18.62 illustrates a simple method whereby a PMOS device sensing the noise between the on-chip supply and ground lines injects a current into an external 50-Ω transmission line and measurement apparatus [2]. Since the transconductance of M_1 can be determined by a small, static change in V_{DD}, the measurement readily reveals both the magnitude and the shape of the supply noise.

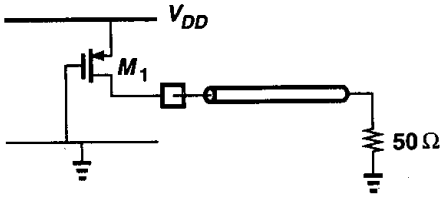


Figure 18.62 Measurement of supply noise.

In cases where a single connection to the chip sustains a prohibitively large transient voltage (e.g., if in Fig. 18.60 or 18.61 many inverters switch simultaneously), multiple pads, bond wires, and package pins are used, decreasing the equivalent inductance (Fig. 18.63).

Example 18.7

In a 600-MHz, 2-V CMOS microprocessor containing 15 million transistors, the supply current varies by 25 A in approximately 5 ns [5]. If the processor provides 200 bond wires for ground and 200 for V_{DD}, estimate the resulting supply bounce.

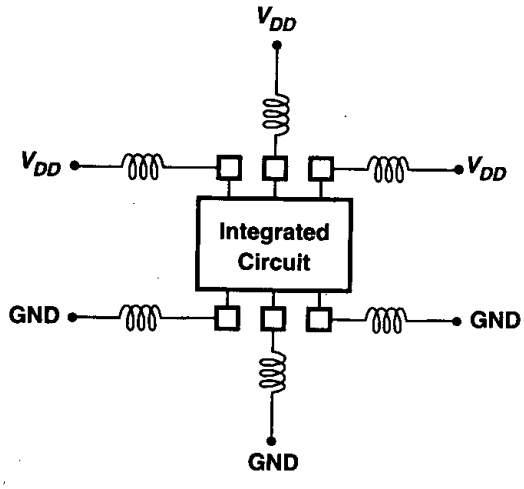


Figure 18.63 Use of multiple wires to reduce overall inductance.

Solution

Assuming a total inductance of 5 nH for each bond wire and its corresponding package trace and pin, we have

$$\Delta V = L\frac{\Delta I}{\Delta t} \tag{18.23}$$

$$= \frac{5 \times 10^{-9}}{200} \cdot \frac{25}{5 \times 10^{-9}} \tag{18.24}$$

$$= 125 \text{ mV} \tag{18.25}$$

In the worst case, the supply bounce and the ground bounce add in-phase, yielding a total noise of roughly 250 mV, greater than 10% of the nominal supply voltage. To further suppress the noise, an external 1-μF MOS capacitor is placed on top of the chip and another 160 supply and ground bond wire pairs are connected from the chip to the capacitor [5].

In some applications, high transient currents drawn from the supply make it difficult to maintain a small bounce on the supply and ground individually. In such cases, a large on-chip capacitor may be used to stabilize the *difference* between V_{DD} and ground. Illustrated in Fig. 18.64, the idea is that if C_1 is sufficiently large, then V_{DD1} and GND_1 bounce in unison. As mentioned earlier, the residual noise on GND_1 may be negligible if the input signals are differential.

This remedy nonetheless involves several issues. First, the value of the capacitor must be chosen carefully because it may otherwise *resonate* with the package inductance at the operating frequency of the chip (e.g., the clock frequency or its harmonics or subharmonics), thereby *amplifying* the supply and ground noise. For this reason, some resistance is added in series with the capacitor (or a MOS capacitor is sized such that its channel resistance dampens the resonance) [5]. Even in the absence of exact resonance, an insufficient value of the decoupling capacitor may simply give rise to slower ringing on the power lines. Second,

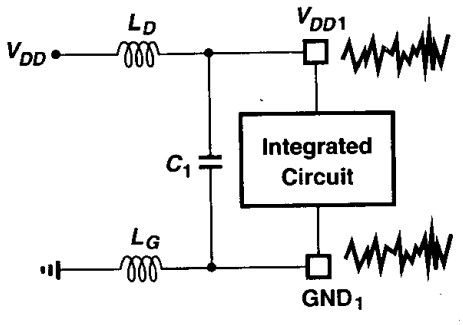


Figure 18.64 On-chip capacitor used to lower supply-ground noise voltage.

since the capacitor is usually formed by a very large MOS transistor (actually, as explained in Section 17.7.2 , a large number of MOSFETs in parallel), the yield of the circuit may suffer. This is because, for the capacitor to be effective, its total area is typically comparable with the total gate area of all of the transistors in the circuit, e.g., it is as if the number of transistors on the chip were doubled.

Self-inductance also manifests itself in the connection to the substrate. As mentioned in Section 18.3, with the large transient currents injected by the devices into the substrate, a low-impedance connection is necessary to minimize the substrate bounce. As shown in Fig. 18.65, some modern packages contain a metal ground plane to which the die can be attached by conductive epoxy. The plane ends in several package pins that are tied to the board ground. Avoiding bond wires and long, narrow traces in the substrate connection, such packages substantially reduce the substrate noise with no additional assembly cost. In more expensive packages, the ground plane is exposed on the bottom and can be directly attached to the board ground, thus avoiding the inductance of the package pins. Also, the ground pads of the circuit can be "downbonded" to the underlying plane to minimize their inductance (while increasing the cost).

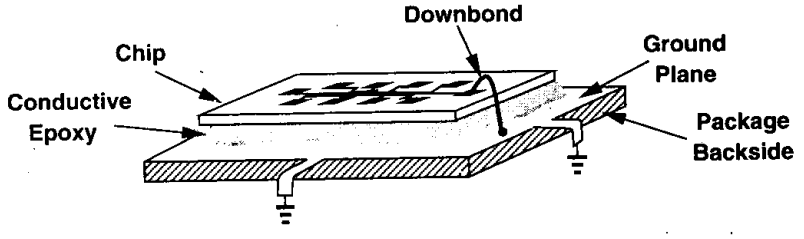


Figure 18.65 Package using a ground plane for substrate connection.

The effect of self-inductance must also be considered for input signals. The inductance along with the pad capacitance and the circuit's input capacitance forms a low-pass filter, attenuating high-frequency components and/or creating severe ringing in transient waveforms. For example, in the precision multiply-by-two circuit described in Section 12.3.3, when the two capacitors are switched to the input, package inductance may limit the settling speed.

Some ICs require constant voltages that must be provided externally. Such voltages may serve as an accurate reference, e.g., in A/D or D/A converters, or to define some bias points on the chip. The package inductance degrades the settling behavior if the circuit injects significant switching noise into the reference.

Example 18.8

Differential pairs are often used as "current switches." As shown in Fig. 18.66, the circuit routes its tail current to either of the outputs according to the large swings controlling the gates of M_1 and M_2. Explain what happens at node X during switching. If the tail currents of a large number of differential pairs feed from node X, should this voltage be provided externally?

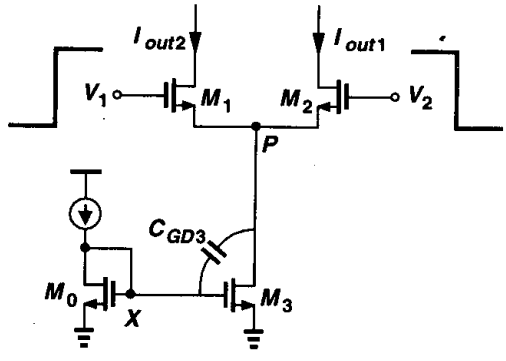


Figure 18.66 Differential pair operating as a current switch.

Solution

Recall from Chapter 4 that for the differential pair to experience complete switching, the differential swing $|V_2 - V_1|$ must exceed $\sqrt{2}(V_{GS} - V_{TH})_{eq}$, where $(V_{GS} - V_{TH})_{eq}$ is the overdrive of M_1 and M_2 in equilibrium, i.e., when $I_{D1} = I_{D2}$. We denote the voltage at node P when the pair is completely switched by V_{P1}, and in equilibrium by V_{P2}. Thus,

$$V_{P1} = V_2 - \sqrt{2}(V_{GS} - V_{TH})_{eq}. \tag{18.26}$$

In equilibrium,

$$V_{P2} = \frac{V_1 + V_2}{2} - (V_{GS} - V_{TH})_{eq}. \tag{18.27}$$

Assuming $V_2 - V_1 = \sqrt{2}(V_{GS} - V_{TH})_{eq}$ and hence $V_1 = V_2 - \sqrt{2}(V_{GS} - V_{TH})_{eq}$, we have

$$V_{P2} = V_2 - \left(1 + \frac{\sqrt{2}}{2}\right)(V_{GS} - V_{TH})_{eq}. \tag{18.28}$$

Thus, V_{P2} is *lower* than V_{P1} by $(1 - \sqrt{2}/2)(V_{GS} - V_{TH})_{eq}$, indicating that during switching V_P drops by this amount. This voltage change is coupled to node X through the gate-drain overlap capacitance of M_3, disturbing I_{D3} and hence I_{out1} or I_{out2}.

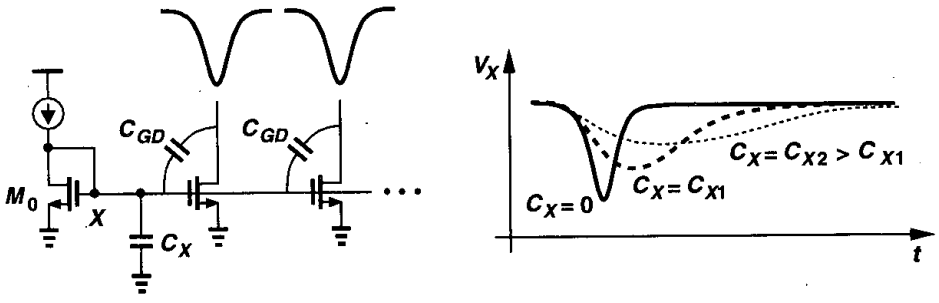


Figure 18.67 Addition of on-chip bypass capacitor to suppress noise at node X.

With a large number of current switches connected to node X, the disturbance may be quite significant, demanding that a decoupling capacitor be connected from node X to ground (Fig. 18.67). However, such a capacitor along with the small-signal resistance of M_0 introduces a long settling time at node X, possibly degrading the overall speed. To avoid this effect, C_X may need to be 100 to 1,000 times the *total* gate-drain overlap capacitance that injects noise into X. If such a large capacitor is placed off-chip, it actually appears in series with the package inductance (Fig. 18.68). In general, careful simulations are necessary to determine the preferable choice here. In many cases, leaving node X agile yields the fastest settling.

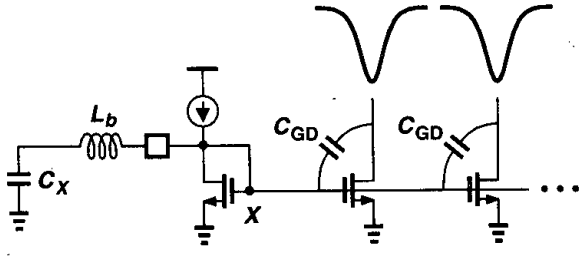


Figure 18.68 Addition of bypass capacitor externally.

The self-inductance of package connections also impacts the performance of digital output buffers. In high-speed systems, these drivers must deliver tens of milliamps of current to the load with fast transitions. With many such buffers operating in a mixed-signal circuit, the resulting voltage drops on the power lines may become very large, increasing the risetime and falltime of the digital outputs and corrupting their timing.

Mutual Inductance While dedicating separate power lines to analog and digital sections reduces the noise on the analog supply, some noise may still couple to sensitive signals through the mutual inductance of bond wires and package traces. As illustrated in Fig. 18.69, both analog supplies and analog inputs are susceptible to noise or transitions on digital supplies, clock lines, or output buffers. With an arbitrary pad configuration, even differential signaling cannot eliminate this effect because the noisy lines may not surround the sensitive lines symmetrically. Thus, the design of the pad frame and the position of the pads play a critical role in the performance that can be achieved.

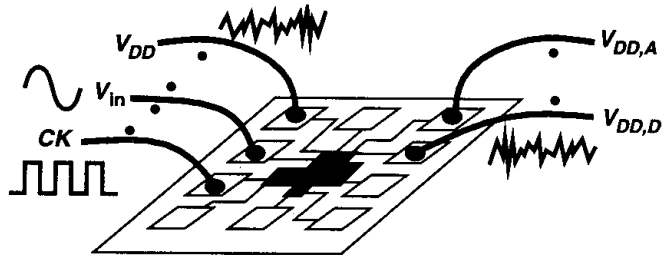


Figure 18.69 Coupling due to mutual inductance between wires.

Mutual inductance also manifests itself in parallel bond wires used to lower the overall self-inductance of a connection (Fig. 18.70). For two such wires, the equivalent inductance is equal to $(L_S + M)/2$, where M denotes the mutual inductance, rather than $L_S/2$.

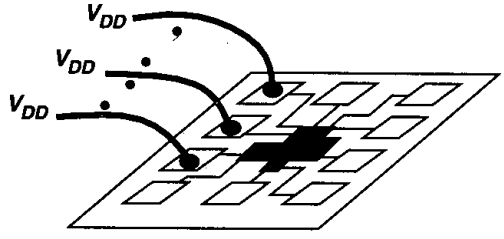


Figure 18.70 Multiple supply bond wires with mutual coupling.

Two methods can reduce the mutual coupling between inductors. First, the wires can be connected such that they are perpendicular to each other, i.e., they terminate on perpendicular sides of the chip [Fig. 18.71(a)]. Second, (quiet) ground or supply lines can be interposed between critical bond wires [Fig. 18.71(b)]. As shown in Fig. 18.71(c), even if several parallel lines are surrounded by ground wires, the effect of mutual inductance drops to negligible values.

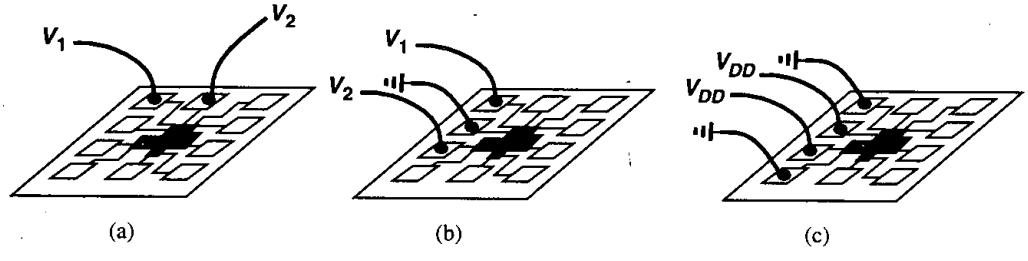


Figure 18.71 Reduction of mutual coupling by (a) perpendicular lines, (b) additional ground lines, (c) occasional ground lines.

It is also interesting to note that mutual inductance *decreases* the self-inductance of two wires if they carry currents in opposite directions. If, as shown in Fig. 18.72, the supply and

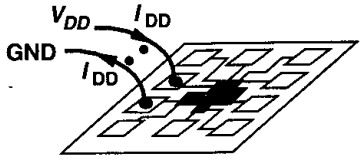


Figure 18.72 Reduction of mutual inductance between two wires carrying equal and opposite currents.

ground lines of a circuit are in parallel, then the total inductance equals $2L_S - M$ rather than $2L_S$. This observation proves useful in designing the pad frame and determining the package connections.

Self- and Mutual Capacitance The capacitance seen from each trace of the package to ground may limit the input bandwidth of the circuit or load the preceding stage. More importantly, this capacitance and the total inductance of the bond wire and the package trace yield a finite resonance frequency that may be stimulated by various transient currents drawn by the circuit. Since the wires and traces exhibit a small series resistance, a high quality factor (Q) results, giving rise to a sharp resonance and amplifying the noise considerably. The capacitance between the traces leads to additional coupling between lines and must be included in simulations.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume $V_{DD} = 3$ V where necessary. Also, assume all transistors are in saturation.

18.1. In Fig. 18.3, polysilicon has a sheet resistance of 30 $\Omega/\square$ (before silicidation) and metal 1 a sheet resistance of 80 $m\Omega/\square$. What is the ratio of the resistivities of the two materials?

18.2. A MOSFET with $W/L = 100\ \mu m/0.5\ \mu m$ undergoes ideal scaling by a factor of two. What happens to the sheet resistivity and the total resistance of the gate?

18.3. A cascode structure uses $W/L = 100\ \mu m/0.5\ \mu m$ for both the input device and the cascode device. If the sheet resistance of polysilicon is 5 $\Omega/\square$ and the maximum tolerable gate resistance 10 Ω, draw the layout of the structure while minimizing the drain junction capacitances.

18.4. In Fig. 18.7, explain what happens to the differential amplifier if each of the design rules A_1-A_8 is violated.

18.5. The input differential pair of an amplifier is to be laid out as in Fig. 18.19 but with each half device (e.g., $1/2M_1$) using four gate fingers. What is the minimum number of interconnect layers required here?

18.6. Large integrated circuits may suffer from significant temperature gradients. Compare the performance of the circuits shown in Fig. 18.21 and 18.22 in such an environment.

18.7. Suppose polysilicon with silicide block has a sheet resistance of 60 $\Omega/\square$ and a parallel-plate capacitance of 100 $aF/\mu m^2$ to the substrate. Also, assume that these parameters are respectively equal to 2 $k\Omega/\square$ and 1000 $aF/\mu m^2$ for the n-well. Determine which material should be used to construct a 500-Ω resistor if matching considerations require a minimum poly width of 3 μm and a minimum n-well length of 6 μm. Neglect fringe capacitances.

18.8. Using the data in Table 17.1, calculate C and C_P for each structure in Fig. 18.33 and identify the one with minimum C_P/C. Neglect fringe capacitances.

18.9. A metal 4 wire with a length of 1000 μm and width of 1 μm is driven by a source impedance of 500 Ω. Using the data in Table 17.1 and assuming a sheet resistance of 40 m$\Omega/\square$, calculate the delay through the wire and compare the result with the lumped time constant obtained by multiplying the source impedance by the total wire capacitance.

18.10. Repeat Problem 18.9 if the width of the wire is increased to 2 μm.

18.11. An interconnect having a length of 1000 μm is required in a circuit. Using the data in Table 17.1 and assuming that the sheet resistance of metal 1–3 is 80 m$\Omega/\square$ and that of metal 4 is 40 m$\Omega/\square$, determine which metal layer must be used to obtain the minimum delay.

18.12. Some new technologies use copper for interconnects because its resistivity is about half that of aluminum. Repeat Problem 18.11 with copper interconnects.

18.13. In the circuit of Fig. 18.50(a), $(W/L)_1 = 100/0.5$ and $I_{D1} = 1$ mA. If the substrate noise, V_{sub}, has a peak-to-peak amplitude of 50 mV, what is the effect referred to the gate of M_1?

18.14. Suppose two bond wires are placed 5 mm above ground with a center-to-center spacing of 1 mm.

(a) What is the total mutual inductance if each wire is 4 mm long?

(b) If one wire carries a 100-MHz sinusoidal current with a peak amplitude of 1 mA, what is the voltage induced across the other wire?

18.15. In Problem 18.14**b,** what center-to-center spacing is required to decrease the induced voltage by a factor of four?

18.16. In order to reduce the total bond wire inductance, a package uses 4 supply pads and 4 ground pads. Suppose the self-inductance of each wire is 4 nH and the mutual inductance between adjacent lines 2 nH. Neglecting mutual inductance between nonadjacent lines, calculate the equivalent inductance of the supply and ground connections if (a) all of the supply wires are placed next to each other and so are the ground wires, (b) every supply wire is placed next to a ground wire.

18.17. The input bandwidth of high-speed circuits may be limited by the bond wire inductance and the pad capacitance. Consider two cases: (a) the bond wire diameter is 50 μm and the pad size 100 μm $\times$ 100 μm; (b) the bond wire diameter is 25 μm and the pad size 50 μm $\times$ 50 μm. If all other dimensions are constant, which case is preferable?

References

1. N. C. C. Lu et al., "Modeling and Optimization of Monolithic Polycrystalline Silicon Resistors," *IEEE Trans. Electron Devices,* vol. ED-28, pp. 818–830, July 1981.
2. D. Su et al., "Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrated Circuits," *IEEE J. of Solid-State Circuits,* vol. 28, pp. 420–430, Apr. 1993.
3. T. Blalack and B. A. Wooley, "The Effects of Switching Noise on an Oversampling A/D Converter," *ISSCC Dig. of Tech. Papers,* pp. 200–201, Feb. 1995.
4. B. Razavi, *Principles of Data Conversion System Design,* New York: IEEE Press, 1995.
5. D. W. Dobberpuhl, "Circuits and Technology for Digital's StrongARM and ALPHA Microprocessors," *Proc. of 17th Conference on Advanced Research in VLSI,* pp. 2–11, Sept. 1997.
6. N. K. Verghese, T. J. Schmerbeck, and D. J. Allstot, *Simulation Techniques and Solutions for Mixed-Signal Coupling in Integrated Circuits,* Boston: Kluwer Academic Publishers, 1995.