Understanding Wide-band MOS Transistors

Fine-line MOS transistors are now fast enough to compete with bipolar detectors and amplifiers, but quantitative understanding of MOS device speed has not kept pace.

High-frequency analog MOS circuits that amplify and detect signals well above audio frequencies have appeared in the last few years [1, 2, 3]. These circuits demonstrate that fine-line MOS transistors are now sufficiently fast to perform in what has been the bipolar transistor's domain. But a quantitative understanding of MOS-transistor speed has been slow to emerge. For example, given a particular CMOS process with a minimum channel length of 2 μm, what amplifier bandwidth can we achieve?

This lack of understanding stems from the absence a commonly-agreed-upon figure of merit for MOS-transistor speed and a lack of familiarity among designers with MOS-amplifier topologies. These problems can be addressed through the use of f_T for MOS transistors, the use of f_T in the prediction of amplifier bandwidth, and a wider familiarity among designers with practical examples of MOS wide-band amplifiers.

Figuring merit

When a new MOS process is developed, the first test circuit is almost always a ring-oscillator chain. This is appropriate because digital circuits drive MOS process development, and ring oscillators give a quick indication of the minimum gate delay for a process. Inferring analog performance from ring-oscillator speed is not so straightforward because information about the ring's topology and operating point are usually not available.

A useful stand-alone figure of merit for transistors in the unity-gain current frequency (f_T). This figure of merit makes intuitive sense for current-controlled bipolar device, but is it a useful measure of an MOS transistor's capabilities? In fact, it is. Although a

by John M. Steininger

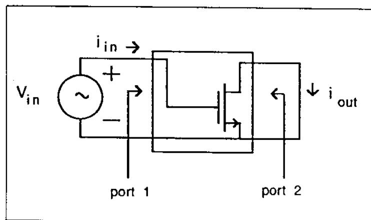
MOSFET's gate draws no current at DC, the displacement current through the gate-to-source capacitance becomes the primary limitation to the transistor's response at high frequency. We can better understand how the f_T of an MOS transistor can be calculated and measured by viewing the transistor as a two-port device (Fig. 1). This two-port's small-signal, linear response can be characterized by any complete set of two-port parameters, such as h-parameters.

The f_T of a two-port is defined as the frequency where the magnitude of h_{21} is unity—where $|i_{out}/i_{in}| = 1$. We can also calculate the f_T of an MOS transistor using the equivalent circuit elements of the hybrid-π model (Fig. 2a), which yields:

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_{gs} + C_{gd}} = \frac{1}{2\pi}\frac{g_m}{C_g};$$

$$\text{where } C_g = C_{gs} + C_{gd}$$

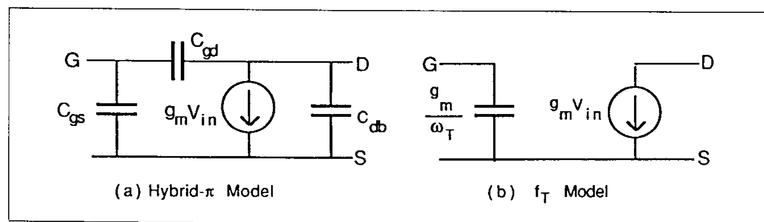
Typically, an n-channel MOS transistor fabricated in a 2μ CMOS process with $W/L = 300/2$, $V_{ds} = 3$ V, and



1. An equivalent circuit of an MOS transistor as a two-port device.

$I_D = 3$ mA will have an f_T of 2 GHz. Since $\omega_T = g_m/C_g$, we can express the gate capacitance of an MOS transistor as $C_g = g_m/\omega_T$. And this allows us to put forth a simple f_T model for an MOS transistor (Fig. 2b).

This f_T model is useful if the drain-to-bulk capacitance, C_{db}, is small compared to C_g, and if the transistor is used



2. Small-signal MOS transistor models.

in a low-gain, wide-bandwidth configuration where the Miller-multiplied C_{gd} can be neglected.

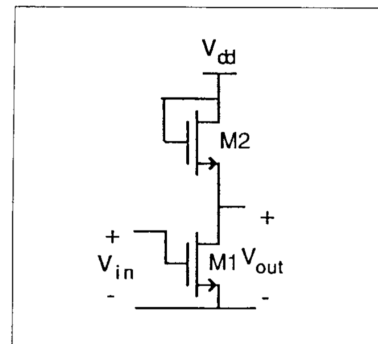
What bandwidths are possible?

Wide-band amplifier configurations in MOS have a different look than the resistively loaded gain stages found in bipolar amplifier designs. The MOS versions are often resistorless because resistors, although available in most modern CMOS technologies, have been neglected by process developers. This neglectful attitude has been shared by analog MOS designers, who have sought to deviate as little as possible from all-transistor, "digital-like" circuit structures. The simplest "all-transistor" realization of a wide-band MOS amplifier is the enhancement-mode inverter (Fig. 3). Here the two active devices are biased in the region where both are active, so $V_{in} = V_{out}$. This amplifier's gain is the ratio of the transconductance of the two transistors. Since the current through the two devices is identical, the gain is set by the ratio of the device sizes:

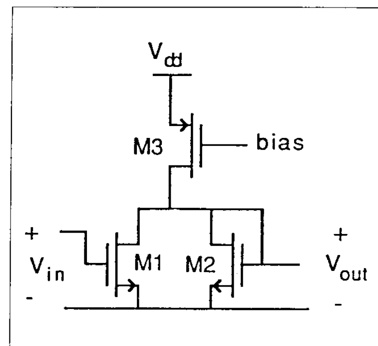
$$G = \frac{g_{m1}}{g_{m2}} = -\sqrt{\frac{Z_1/L_1}{Z_2/L_2}}$$

Unfortunately, the current through this amplifier is not well controlled and the output swing can go no higher than $V_{dd} - V_{thn}$.

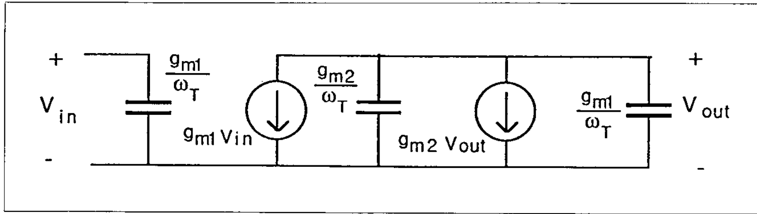
It is possible to design a more practical amplifier in which the gain is still set by the ratio of device geometries (Fig. 4). Here, the bias current is set by current source M3 and the output is capable of a wider swing. We can calculate the achievable bandwidth of this



3. Enhancement-mode inverter.



4. Simple practical CMOS amplifier.



5. A small-signal equivalent circuit of the amplifier shown in Fig. 4 based on the f_T transistor model.

simple gain stage by using the f_T transistor model to analyze the small-signal equivalent circuit (Fig. 5).

If we neglect the capacitance loading of M3 and assume the circuit is loaded by an identical following stage, the stage gain is $A_0 = -g_{m1}/g_{m2}$. The 3 db bandwidth is:

$$\omega_p = \frac{g_{m2}}{C_{total}} = \frac{g_{m2}}{\frac{g_{m1}}{\omega_T} + \frac{g_{m2}}{\omega_T}} = \frac{\omega_T}{A_0 + 1}$$

and $f_p/f_T = 1/(A_0 + 1)$.

To obtain an amplifier with the highest possible bandwidth, we would like to cascade stages of low gain. (See sidebar 1.) For a stage gain of $A_0 = 3$, we can calculate that the maximum attainable bandwidth of an amplifier made in a 2μ CMOS technology with $f_T = 2$ GHz is $f_{3db\,max} = 500$ MHz.

Scaled MOS and f_T

We can increase the f_T of an MOS device by making the g_m larger or the C_g smaller, or both. A simple, first-order theory that relates f_T to device properties can be developed if we approximate the gate capacitance of a MOS transistor in a way that neglects the gate-to-drain overlap capacitance. If we assume that C_g is only the parallel plate capacitance between the gate and channel, we can then say that $C_g \approx C_{ox}WL$. We know that:

$$g_m = \frac{d}{dV_{gs}} I_d$$

$$= \frac{d}{dV_{gs}}\left[\frac{1}{2}\mu_{eff}C_{ox}\frac{W}{L}(V_{gs} - V_{th})^2\right]$$

$$= \mu_{eff}C_{ox}\frac{W}{L}(V_{gs} - V_{th})$$

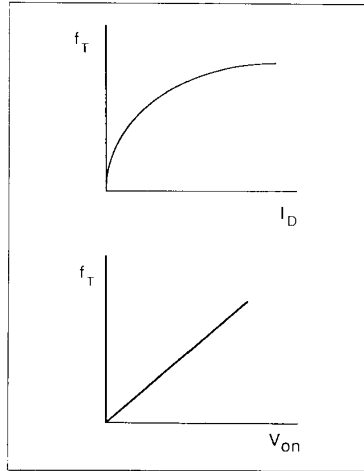
$$= \mu_{eff}C_{ox}\frac{W}{L}V_{on}, \quad \textit{where}$$

$$V_{on} = V_{gs} - V_{th}$$

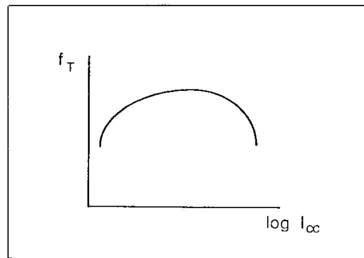
From these two equations we can write a simple expression for f_T:

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_g} \approx \frac{1}{2\pi}\frac{\mu_{eff}V_{on}}{L^2}$$

This simple expression has two important implications. First, f_T is proportional to $1/L^2$. Since process technologies tend to scale down by a factor of $1/\sqrt{2}$ per generation, each new technology should have twice the f_T of the old, provided that the devices continue to exhibit square-law behavior. Second, in a given technology, devices made with the minimum channel length and biased with the largest V_{on} will have the greatest f_T, regardless of W.



6. f_T vs. V_{on} and I_D.



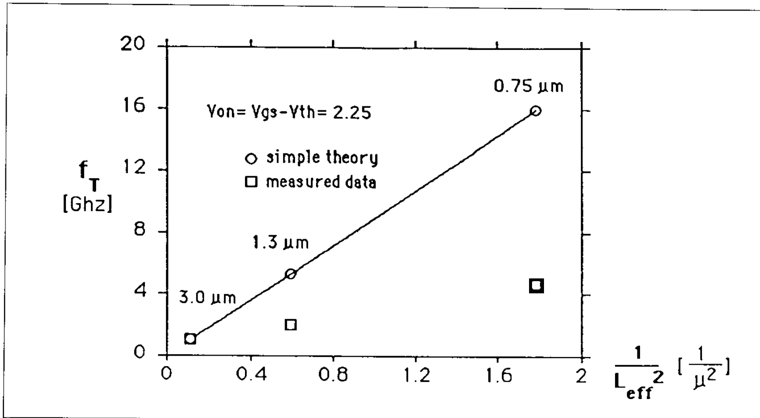
7. f_T of a bipolar transistor vs. log I_{CC}.

We can see from the above equations that f_T is an unbounded increasing function of both V_{on} and I_D (Fig. 6). This differs from the situation in a bipolar transistor, where f_T peaks at some optimum collector current (Fig. 7). This roll-off in bipolar f_T at high currents is due to an increase in τ_F caused by high-level injection and the Kirk effect, the same mechanisms that cause β_F to drop off [4].

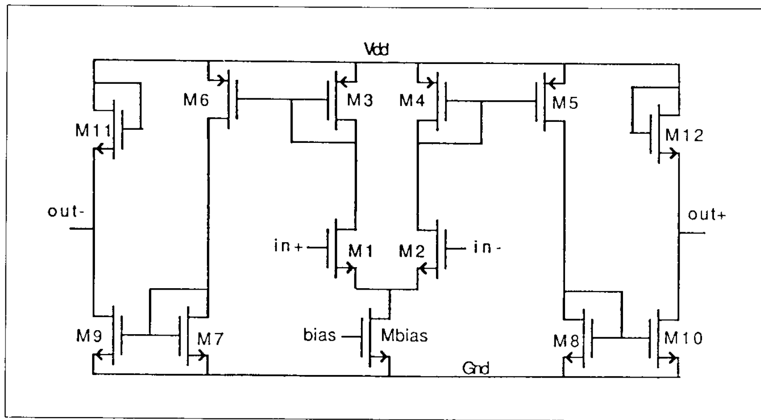
It is natural to compare the f_T of bipolar transistors at their optimum collector current. But no such f_T peak occurs in MOS, so at what point should MOS transistors from different technologies be compared? That point should be some practical value of V_{on} that is held constant across technologies. A value of V_{on} that is commonly used to measure the drive-current capability of MOS transistors for digital applications ($V_{supply} - V_{th}$) is not appropriate for analog design because it is not a practical biasing point.

In a 5-V technology, the upper limit of a practical bias is approximately $V_{on} = 2.25$ V, and this is the point at which we have plotted measurements of f_T vs. L_{eff} for 3 CMOS technologies (Fig. 8). Also plotted are the calculated f_T's of these processes using the simple f_T theory and normalized to the 3μm CMOS measurement. The measured f_T's do not follow the simple theory in which f_T is proportional to $1/L^2$. Instead, the increase in f_T with decreasing gate length is shown to be much less dramatic, and there are two reasons for this. First, short-channel transistors operated at practical bias voltages do not behave as square-law devices. Their short channel lengths and correspondingly thin gate dielectrics (<500 Å) produce high electric fields in the channel, both in the longitudinal and transverse directions. These high fields degrade the drift velocity of the channel carriers—often called velocity saturation—which produces a smaller-than-expected transconductance for these devices [5, 7].

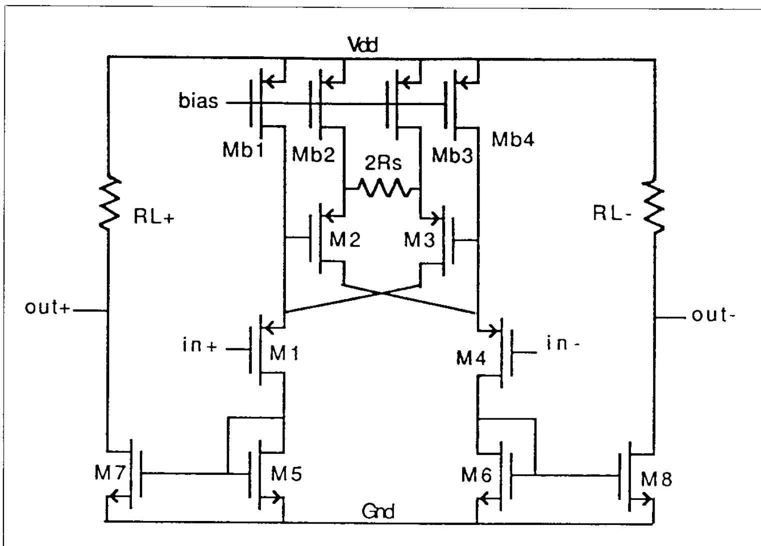
Second, the simple theory neglects the gate-to-drain overlap and assumes



8. f_T vs. L_{eff} for various CMOS technologies.



9. Simple differential wideband amplifier.



10. Resistively loaded MOS amplifier with g_m cancellation.

that C_g is only a parallel-plate capacitance even though small gate geometries have fringing fields that contribute significantly to C_g. These field lines to not terminate on charge in the channel and therefore do not contribute to the transconductance of the device.

These two factors produce a measured f_T for scaled MOS technologies that is lower than that predicted by first-order theory. Still, on a absolute scale, the speed of these transistors is quite respectable, and CMOS circuits with wide bandwidth are possible.

Wide-band CMOS amplifiers

Although the simple-wide-band CMOS amplifier shown in Fig. 4 is practical, its application is limited by its single-ended nature. Differential wide-band amplifiers are more versatile, due to their floating common-mode input range, increased output swing, and the first-order cancellation of nonlinearities in their differential outputs. Another advantage of differential circuits is their potential for increased power-supply rejection. This is particularly important at high frequencies where active power-supply rejection is not possible.

A differential amplifier where gain is again set by the ratio of g_m's is shown in Fig. 9. Here, current mirrors are used to provide additional current gain. The voltage gain of this amplifier is:

$$G = \frac{g_{m1}(W/L)_{M6}(W/L)_{M9}}{g_{m11}(W/L)_{M3}(W/L)_{M7}}$$

A limitation of this scheme is that dc bias current, as well as signal current, is amplified by the current mirrors. This can be overcome by using interstage dc subtractors [1]. If this and other limitations—the inherent non-linearity of the single-ended outputs and the limited output swing—can be tolerated, the circuit can provide very wide bandwidth.

For applications where wide output swing and good single-ended linearity are most important, being able to incorporate resistors in the design would be helpful. But combining resistors and MOSFETs in amplifiers is problematic, primarily because of the g_m of MOSFET transistors. To understand this, consider a bipolar and a MOS transistor biased at the same current and dimensioned so that they have the same f_T. The g_m of

Optimum Stage Gain for Maximum Cascaded Bandwidth

It is possible to show that there is an optimum stage gain for a cascade of identical amplifiers that maximizes the overall bandwidth. Consider a cascade of several amplifiers with an equivalent circuit like that of Fig. 5. The frequency transfer function of this circuit will be:

$$A(s) = \left[\frac{A_0}{1 + \frac{s}{\omega_p}}\right]^n \quad \text{(a1.1)}$$

where A_0 and ˇp are the gain and 3db bandwidth of an individual stage. We can solve Equation a1.1 for h, the 3db frequency of the overall cascade, such that:

$$\frac{|A(s)|^2}{A_0^{2n}} = \frac{1}{\left[1 + \left(\frac{\omega_h}{\omega_p}\right)^2\right]^n} = \frac{1}{2}$$

This yields:

$$\omega_h = \omega_p\sqrt{2^{1/n} - 1} \quad \text{(a1.2)}$$

We know that:

$$\omega_p = \frac{g_{m2}}{C_{\text{total}}} = \frac{g_{m2}}{\frac{g_{m1}}{\omega_T} + \frac{g_{m2}}{\omega_T}} = \frac{\omega_T}{A_0 + 1}$$

The total cascaded gain will be the product of each of the individual gains, A = Aon. Substituting these into Equation a1.2 gives:

$$\frac{\omega_h}{\omega_T} = \frac{\sqrt{2^{1/n} - 1}}{A^{1/n} + 1} \quad \text{(a1.3)}$$

Equation a1.3 has a maxima that is independent of A. We can find that maxima analytically if we make a couple of assumptions. First, for large n ($n > 4$), we can say that:

$$\omega_h = \omega_p\sqrt{2^{1/n} - 1} \approx \omega_p\frac{0.83}{\sqrt{n}} \quad \text{(a1.4)}$$

to a good approximation. Second, assume that the capacitance of the load conductance can be neglected, that is g_{m1}/ˇT ≫ g_{m2}/ˇT. Then:

$$\omega_p \approx \omega_T\frac{g_{m1}}{g_{m2}} = \frac{\omega_T}{A_0} = \frac{\omega_T}{A^{1/n}}$$

Combining this with Equation a1.4 gives a new expression for a1.3:

$$\frac{\omega_h}{\omega_T} = \frac{0.83}{A^{1/n}\sqrt{n}} \quad \text{(a1.5)}$$

the MOS transistor, even in a fine-line technology, will typically be an order of magnitude less than the g_m of the bipolar transistor. If we construct a simple one-transistor amplifier, using R_L as the load resistor and R_S as the degeneration resistor in series with the source or emitter, the gain will be:

$$G = \frac{R_L}{R_S + \frac{1}{g_m}}$$

In the bipolar case, the gain of this amplifier can be made independent of g_m by choosing $R_s \gg 1/g_m$. To meet this same criterion in the MOS case, R_L and R_S must be at least an order of magnitude larger. This becomes a problem if bias current passes through these resistors because their voltage drop can quickly exceed the power supply's available head room. This suggests that MOS amplifier designs that include resistors should ensure that bias current does not flow through the gain-setting resistors.

Large-valued resistors take up area and have correspondingly large parasitic capacitances that limit amplifier bandwidth. It would be more convenient if we could use smaller resistors in our amplifier designs and still have a gain that is independent of g_m. In fact, we can design a circuit that has its gain set by a ratio of resistors and removes the effect of the MOSFET's g_m on the circuit gain by using a cancellation scheme (Fig. 10).

This cancellation is accomplished by cross coupling the drains of M2 and M3 to the sources of M1 and M4. This insures that $V_{gs1} + V_{gs2} = V_{gs3} + V_{gs4}$. If we apply a small differential voltage V_d across the input terminals, it appears directly across the resistor, $2R_S$. The resulting current, $V_d/2R_S$, is added or subtracted from the circuit's bias current and mirrored into the load resistors. The gain of this circuit, provided that $I_{b1,4} \ll I_{b2,3}$, is then:

$$G = \frac{R_L(W/L)_{M7}}{R_S(W/L)_{M5}}$$

It is even possible to build resistively loaded MOS amplifiers with no source degeneration at all and still have the gain set by a ratio of resistors (Fig. 11). This differential amplifier uses a tapped resistive load and a simple output common-mode setting network. By cen-

Derivation of the Bias Current that Produces $g_m = 1/R_b$

The bias circuit shown in Fig. 12 produces a current that will set the gm of a matched transistor equal to $1/R_b$. Referring to Fig. 12, if we sum the voltages around the loop formed by MB2, MB1, and R_b, we get:

$$V_{gs2} = V_{gs1} + I_bR_b$$

Since:

$$V_{gs} = \sqrt{\frac{I_d}{K}} - V_{th}:$$

$$\text{where } K = \frac{\mu_{eff}C_{ox}}{2}\frac{W}{L}$$

Then:

$$\sqrt{\frac{I_b}{K}} = \frac{1}{2}\sqrt{\frac{I_b}{K}} + I_bR_b$$

Solving for I_b yields:

$$I_b = \frac{1}{4KR_b^2} = \frac{1}{2\mu_{eff}C_{ox}(W/L)R_b^2}$$

The g_m of a device with dimensions (W/L) and biased at I_b is:

$$g_m = \sqrt{2\mu_{eff}C_{ox}(W/L)I_b} = \frac{1}{R_b}$$

We desire to find the maxima of this function. To do so, we can take the derivative of the natural logarithm of both sides—maximizing the logarithm will maximize the expression—and set equal it equal to zero:

$$\frac{d}{dn}\left(\ln\frac{\omega_h}{\omega_T}\right) = \frac{\ln A}{n^2} - \frac{1}{2n} = 0$$

Now, solving for the stage gain gives the very simple final result:

$$A_0 = A_1/n$$

Thus, the maximum overall bandwidth can be achieved with cascaded stages of very low gain. This relation is exact only for the amplifier configuration specified. A plot of a1.3 and a1.5 vs. stage gain for a fixed A ($A = 50$) is shown in Fig. A_1.

Notice that the plot of a1.3, which represents the f_T equivalent circuit, has a broad maxima. Typically, the more complicated the gain stage, the smaller the achievable bandwidth and the broader the peak. The actual choice of stage gain becomes a trade-off between power consumption and ultimate bandwidth.

ter tapping the load resistor and level-shifting through source follower M7, the output common-mode voltage is set at $V_{gs5} + V_{gs7}$. With no source degeneration resistor R_S, the gain of this amplifier is simply $G = g_{m1,2}R_L$.

The cleverness that makes this circuit useful is in the bias circuit, which biases M9 to produce a current that sets the g_m's of M1 and M2 proportional to R_L (Fig. 12). This circuit is an MOS version of a self-biasing Widlar current source [6]. In the bipolar version, a current I_b proportional to the thermal voltage $V_T = kT/q$ is produced. Substituting MOS transistors produces a current $I_b = 1/(4KR_b^2)$ where $K = (\mu C_{ox}/2)(W/L)$. (See sidebar 2)

If the devices are sized such that $(w/L)_{B2} = (W/L)_9 = 2(W/L)_{1,2}$ and $(W/L)_{B3,4} = (W/L)_3 = (W/L)_4$, then a current will bias the input pair M1 and M2 such that $g_{m1,2} = 1/R_b$. Therefore, $G = g_{m1,2}R_L = (4KI_B)^{1/2}R_L = R_L/R_b$, and the gain has been made independent of the MOSFET's g_m and proportional to a ratio of resistors.

As we've seen, scaled MOS technologies have native transistors with high f_T's that can be incorporated into wide-band designs. Although the speed of these fine-line devices does not scale up as rapidly as simple theory would predict, MOSFETs with gate dimensions on the order of 1 μm have f_T's approaching 5 GHz. The f_T figure of merit can be used to predict the performance of simple amplifiers, although practical amplifiers typically have less bandwidth. Practical differential amplifiers that have gains set by ratios of transistors or resistors can be built, and the exhibit gain-bandwidth products approaching the predicted limit of f_T.

11. A MOSFET amplifier with $G = g_mR_L$.

12. Bias circuit for the amplifier of Figure 11 that produces $g_m = R_b$.

References

1. D.M. Pietruszynski et al., "A 50 Mbit/s CMOS Monolithic Optical Receiver," IEEE J. Solid-State Circuits, Vol. SC-23, No. 6, pp. 1426–33 (Dec. 1988).
2. A.A. Abidi, "Gigahertz Transresistance Amplifiers in Fine Line NMOS," IEEE J. Solid-State Circuits, Vol. SC-19, No. 6, pp. 986–94 (Dec. 1984).
3. Jeff Scott et al., "A 16 Mbit/s Data Detector and Timing Recovery Circuit for Token LAN," ISSCC Digest of Technical Papers, Vol. 32, pp. 150–51 (Feb. 1989).
4. P.R. Gray and R.G. Meyer, Analysis and Design of Analog Integrated Circuits, pp. 31, (Wiley, 1977).
5. C.G. Sodini, P.K. Ko and J.L. Moll, "The Effects of High Fields on MOS Device and Circuit Performance," IEEE Tran. on Electron Devices, Vol. ED-31 (Oct. 1984).
6. P.R. Gray and R.G. Meyer, op. cit., pp. 282–83.
7. K.Y. Toh, P.K. Ko and R.G. Meyer, "An Engineering Model for Short-Channel Devices," IEEE J. Solid-State Circuits, Vol. SC-23, No. 4, pp. 950–58 (August 1988). ***CD***

Acknowledgements

Thanks to Dave Pietruszynski, Jeff Scott, and T.R. Viswanathan for their wide-band-amplifier-circuit contributions. Special thanks to O.G. Petersen for his editorial assistance.

Biography

John M. Steininger is supervisor of an analog IC design group at AT&T Bell Laboratories in Reading, Pennsylvania.