CMOS RF: No Longer an Oxymoron

Thomas H. Lee
Stanford University Center for Integrated Systems
CIS-205 MC4070
Stanford, CA 94305-4070 USA
tomlee@ee.stanford.edu

Abstract–Peak CMOS f_T's are now in excess of 30GHz and double every three years. That raw device speed is supplemented by recently developed passive elements, such as the lateral flux capacitor and the shielded spiral inductor, in which the lossy substrate is made much less relevant without requiring special processing steps.

Device F_{min} is typically under 0.5dB at 1-2GHz, and a better understanding of broadband MOSFET noise has shown how to minimize amplifier noise figure within a specified power budget. Finally, a new understanding of phase noise has shown that satisfaction of previously unappreciated symmetry criteria can suppress greatly (e.g., by factors of 5-10 or more) the upconversion of 1/*f* device noise into close-in phase noise.

I. Introduction

CMOS continues to scale dramatically, making it an attractive alternative to more exotic technologies for many RF applications in the low-GHz frequency range. While CMOS certainly suffers from inferior device physics, it is important to note that such inferiority is irrelevant if adequate performance can be provided at the lowest cost. This paper describes several important developments that have allowed CMOS to cross this threshold of sufficiency for many applications once thought inaccessible to CMOS.

II. Passive Components

Because of its digital origins, CMOS technology has never had a focus on providing good passive components. However, creative exploitation of the rich variety of layers and structures available in standard digital CMOS processes offers a number of ways to improve the characteristics of on-chip passive elements. In particular, it is possible to reduce significantly the severity of substrate loss.

A. Inductors

Planar spiral inductors in silicon technology typically have *Q* values well below 10 in the 1GHz frequency range. Their design is often somewhat of a haphazard affair, so the *Q* achieved in practice typically falls considerably short of even the dismal theoretical limit. A recently developed compact analytical model, however, facilitates structured inductor design by permitting the generation of contours of constant *Q* as a function of layout parameters. Armed with these contours, it is a straightforward exercise to determine both the maximum *Q* attainable, and the layout dimensions that produce it.

An important observation is that, aside from skin effect loss, significant *Q* degradation results from energy coupled into the substrate. Image current losses have a measurable effect, but typically do not dominate, and diminish as the separation between the inductor and substrate increases. As the number of available interconnect layers grows, obtaining adequate separation becomes easier. The effective separation distance can be increased even further by placing alternating wedges or strips of *n*-well and substrate underneath the inductor. The reverse-biased *p*-*n* junctions prevent the flow of image currents near the semiconductor surface, increasing the spacing between the inductor and image currents by an amount equal to the depth of the *n*-well.

The more serious substrate loss mechanism, typically accounting for roughly half of the *Q* degradation, is due simply to the flow of currents into the substrate through the parasitic inductor-to-substrate capacitance. To prevent this current from flowing in the substrate, a layer of interconnect may be dedicated as a grounded shield between the inductor and substrate so that *E*-field lines terminate on the low-loss shield rather than the substrate. To prevent the shield from acting as a shorted secondary turn, slots are cut into it to inhibit the flow of eddy currents in the shield layer:

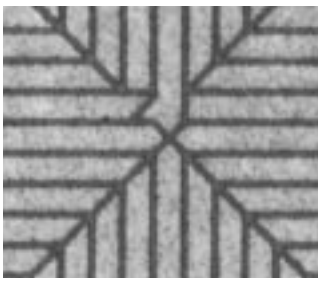

Fig. 1. Patterned ground shield

Such *patterned ground shields* [1] have allowed a near doubling of *Q* of resonators constructed with spiral inductors. Hence, even though planar spirals will never possess extremely high *Q*, these measures mitigate substrate loss to a

degree sufficient to extend greatly their usefulness.

The patterned ground shield also significantly improves isolation by minimizing coupling through the substrate. Experiments reveal that substrate coupling to the inductor may be attenuated by 20dB or more in the low-GHz frequency range over unshielded structures. The important implications of these results for full integration of RF circuitry hardly require expression.

B. Capacitors

Of the many fixed capacitor options available in standard CMOS technologies, gate capacitors provide the highest capacitance per unit area (presently on the order of 5-7fF/μm^2, and increasing with successive process generations). However, linearity suffers if operation in strong inversion is not maintained.

Some applications (e.g., power amplifier matching networks, reactive terminations for mixers, etc.), however, require exceptionally linear capacitors. There, ordinary metal-insulator-metal (MIM) structures are the only practical option. While some processes devoted to analog applications provide a special thinned intermetal dielectric layer, most technologies do not. In those more common cases, MIM capacitors are typically less area efficient than gate capacitors by two orders of magnitude. Furthermore, bottom-plate parasitic capacitance is then frequently about the same value as the main capacitance, leading to numerous well-understood and vexing problems.

Fortunately, scaling trends provide an additional option. The permissible line-to-line spacing in many deep submicron technologies is now smaller than the spacing between interconnect layers. While the substantial adjacent-line capacitance that results is a highly undesirable property for digital interconnect, it offers an opportunity for increasing the areal density of capacitors. By exploiting fully this *lateral* flux, one may build MIM capacitors that require much less area (e.g., by factors of 5-10 or more) than conventional parallel-plate structures, and that are perfectly linear [2]. Furthermore, since lateral flux dominates, the bottom plate's (e.g., substrate's) influence is significantly attenuated. If the upper metal levels are used, the substrate's effects can be eliminated for all practical purposes. Use of an explicit shield layer is also always an option if suppression of substrate effects is particularly important. Additionally, sandwiches of lateral flux capacitors may be used to increase capacitance even further. With five interconnect layers available in many processes, an additional increase in capacitance by perhaps a factor of four is possible through the use of such sandwiches.

Maximum exploitation of lateral flux might result from using geometries based on fractals, since fractals can enclose a finite area with an infinite perimeter. Although photolithographic limitations constrain the capacitance increase to a finite value, quite substantial enhancement is nonetheless possible in practice. The chief, and hopefully temporary, limitation is that extraction tools typically fail quite spectacularly when attempting to analyze a fractal capacitor, so that it is difficult to predict capacitance values.

It is important to underscore that the benefits of lateral flux capacitors, fractal or otherwise, are obtained without requiring any special process modifications. They derive simply from the scaling trends already in place. Since the needs of digital circuits generally dictate CMOS process evolution, these benefits may be expected to improve as scaling continues.

III. Broadband Device Noise

It has been known for quite some time that short-channel MOSFETs in saturation exhibit considerably more broadband RF noise than predicted by long channel theory [3]. This observation has led to speculation that unacceptably poor noise performance might accompany scaling to smaller dimensions.

The thermally noisy channel charge produces effects that are modeled by a drain and gate current noise generator. These currents are partially correlated with each other because they share a common origin, and possess spectral densities given by the following equations:

$$\overline{i_{nd}^2}/\Delta f = 4kT\gamma g_{d0} \tag{1}$$

$$\overline{i_{ng}^2}/\Delta f = 4kT\delta g_g \tag{2}$$

The parameter g_{d0} is the drain-source conductance at zero drain-source voltage, while g_g is

$$g_g = \frac{\omega^2 C_{gs}^2}{5 g_{d0}} \tag{3}$$

In the long-channel limit, values for γ and δ in saturation are 2/3 and 4/3, respectively. As the longitudinal field strength grows, carrier velocities begin to saturate, and further increases in field cause carrier heating and the observed increases in γ and δ. The increase in γ is most rapid once the applied drain voltage exceeds $V_{DS,sat}$, and more gradual once the device is deep in the saturation region:

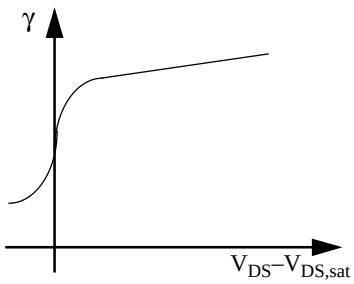

Fig. 2. γ vs. drain-source overdrive (approximate)

Detailed theoretical calculations show that γ is typically about 2-3 times the long-channel value for drain-source voltages that are a few hundred millivolts above $V_{DS,sat}$, and is only a relatively weak function of channel length [4]. Hence, limiting drain-source voltages to avoid deep saturation is an effective way to minimize RF noise degradation. Fortunately, this constraint is in harmony with the decreasing supply voltages that accompany scaling.

While there is reasonable agreement between theory and measurement for γ, there are no published measurements of the behavior of δ. Theoretical considerations suggest that the growth in δ with field could be substantially worse than for γ, but keeping the drain-source overdrive voltage below a few hundred millivolts is also effective here.

Even though short-channel devices suffer from such high-field thermal noise enhancement, it is important to note that scaling improves f_T at the same time, and that the corresponding improvement in device noise figure outpaces degradation due to those high-field effects. Device F_{min} values below 0.5dB at 1GHz are now the norm, and scaling continues to improve device noise performance. Hence, even though other technologies are superior in this regard, the noise figures achievable with CMOS are adequate for many applications.

IV. Power-Constrained LNA Design

While device F_{min} is quite acceptable, there remains the question of how closely amplifier noise figures can approach F_{min} in practice, particularly if there is a constraint on the allowable power consumption. Classical noise optimization methods do not take power consumption explicitly into account, and accommodate constraints on gain, input match and linearity only clumsily, if at all. As a result, traditional low-noise amplifier (LNA) design methods frequently effect unsatisfactory tradeoffs among those parameters. Furthermore, classical two-port noise theory offers little or no guidance about how best to exercise the integrated circuit designer's cherished freedom to select device dimensions.

A narrowband LNA architecture that nicely balances various performance parameters is the inductively degenerated common-source amplifier:

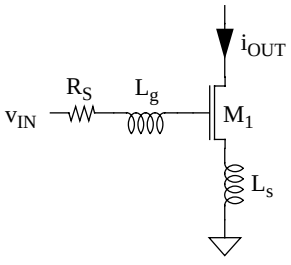

Fig. 3. Inductively-degenerated LNA

The inductance L_S interacts with the gate-source capacitance and device transconductance to produce a resistive component to the input impedance, while the additional gate inductance L_g provides the extra degree of freedom necessary to guarantee operation at resonance. This method of generating a real term to the input impedance is preferable to resistive methods because pure reactances are noiseless. Furthermore, operation at resonance guarantees that good gain is obtained simultaneously with an excellent input match and near-optimum noise figure.

Using approximate analytical device models, it is possible to derive an expression for LNA noise figure with power consumption as a parameter [5]. An insight gained from that exercise is that the minimum noise figure at constant power occurs when the input loop formed by R_S, L_S, L_g, C_{gd} and the effective real part measured at the gate of M_1 has a Q of typically 3-5. This Q fixes, in turn, the optimum width of the transistor at a value given by

$$W_{opt} \approx \frac{1.5}{\omega L C_{ox} R_S Q} \approx \frac{1}{3\omega L C_{ox} R_S} \tag{4}$$

Using values typical of processes now in use, (4) yields the rule-of-thumb that the optimum device width is about 750μm-GHz for a 50Ω system [6]. When this optimum width is used, the minimum power-constrained (MPC) amplifier noise figure is given by:

$$F_{minP} = 1 + 2.4\frac{\gamma}{\alpha}\left[\frac{\omega}{\omega_T}\right] \approx 1 + 5\left[\frac{\omega}{\omega_T}\right] \tag{5}$$

The foregoing derivations assume that the ratio of γ to δ changes little from their long-channel values. Fortunately, the optimum width is sensitive only to the square-root of this ratio.

If ω_T/ω is 10, the MPC noise figure is somewhat better than 2dB, while if ω_T/ω is 20, the MPC noise figure is somewhat better than 1dB. Additional noise sources (e.g., inductor loss, second-stage contributions, etc.) will elevate these figures somewhat, but it remains true that practical noise figures below 2dB are obtainable in the range of 1-2GHz with approximately 10mW of dissipation with production processes, based on laboratory results reported in [2].

Finally, linearity improves with scaling because of the increasing prominence of velocity saturation, which causes g_m to approach a constant value. Input-referred third-order intercept values in excess of 0dbm are routinely achievable for single-stage designs, with values typically degrading to approximately –5 to –10dBm for two-stage designs.

V. Oscillator Phase Noise

Another area in which CMOS has been called deficient is oscillators, primarily because there is a widely held belief that large 1/*f* device noise necessarily implies large close-in phase noise. However, that belief derives from a quasi-heuristic

model of phase noise that assumes that oscillators are time-invariant linear systems. It is easy to show that oscillators are in fact fundamentally periodically time-varying systems, and an explicit acknowledgment of this truth leads to a new understanding of how phase noise evolves from device noise.

Specifically, consider the response of a simple, lossless *LC* oscillating tank to an injected impulse of current. The response depends on the particular phase (modulo 2π) of injection:

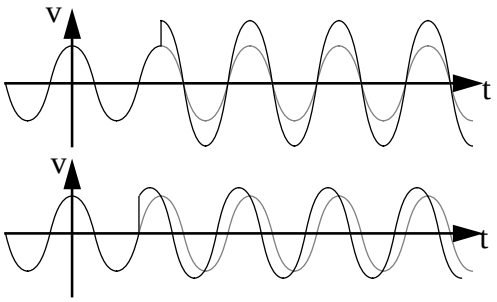

Fig. 4. Waveforms for impulse excitation of *LC* oscillator

If, as in the upper waveform, the impulse happens to coincide with an extremum of the existing oscillation, the net effect is simply a change in amplitude because the impulse response is in phase with the oscillation; the timing of the zero crossings does not change.

On the other hand, if the impulse is injected at some other time, the net effect is generally a change in both amplitude and phase. Note further that the phase change persists for all time. Hence, it is clear that the phase displacement resulting from an impulsive disturbance is a periodically time-varying function, and can be as small as zero if the impulse is injected at just the right instant.

Because an impulse produces a step change in phase, the impulse response may be expressed as follows:

$$h_\phi(t,\tau) = \frac{\Gamma(\omega_0\tau)}{q_{max}} u(t-\tau) \qquad (6)$$

where $\Gamma(x)$ is a periodic function, known as the impulse sensitivity function (ISF), and q_{max} is the maximum charge displacement in the tank. Since $\Gamma(x)$ is periodic, it may be expressed as a Fourier series, and used in a superposition integral to determine the phase noise spectrum resulting from known device and circuit noise.

In a time-varying system, a signal at one frequency can cause a response at other frequencies, and the response at one frequency can be the result of excitation at multiple frequencies. For the specific case of an oscillator, an important insight is that phase noise close to the carrier results from the folding of device noise centered at integer multiples of the carrier frequency:

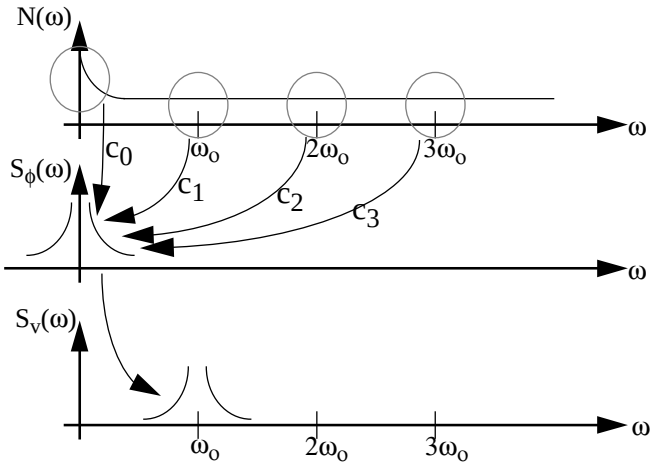

Fig. 5. Evolution of phase noise

As can be seen, the upconversion of device $1/f$ noise occurs through c_0, the DC value of the ISF, and can therefore be suppressed if the ISF has zero DC value. Since the ISF is a function of the actual oscillation waveform, it is under the control of the designer. By satisfying the implied symmetry requirement, then, one can suppress the upconversion of $1/f$ noise into close-in phase noise, as has been experimentally verified [7]. Hence, it is now understood that poor $1/f$ device noise performance does not automatically preclude good phase noise performance.

VI. Summary

It is clear that scaling trends, properly exploited and combined with new insights into device and oscillator noise, enable CMOS IC technology to perform well enough at GHz frequencies to make it attractive for applications once thought the sole province of more exotic technologies.

References

[1] C. Yue, and S. Wong, "On-Chip Spiral Inductors with Patterned Ground Shields for Si-Based RF IC's," *1997 VLSI Circuits Conference Digest of Technical Papers*, June, 1997.

[2] A. Shahani, D. Shaeffer, T. Lee, "A 12mW Wide Dynamic Range CMOS Front-End for a Portable GPS Receiver," *ISSCC Digest of Technical Papers*, Feb. 1997.

[3] A. Abidi, "High-Frequency Noise Measurements on FET's with Small Dimensions," *IEEE Transactions on Electron Devices*, Nov. 1986.

[4] G. Klimovitch, T. Lee, Y. Yamamoto, "Physical Modeling of Enhanced High-Frequency Drain and Gate Current Noise in Short-Channel MOSFETs," *Proceedings of the First International Workshop on Design of Mixed-Mode Integrated Circuits and Applications*, July, 1997.

[5] D. Shaeffer, T. Lee, "A 1.5V, 1.5 GHz CMOS Low-Noise Amplifier," *IEEE Journal of Solid-State Circuits*, May, 1997.

[6] T. Lee, *The Design of CMOS RF Integrated Circuits*, Cambridge University Press, 1998.

[7] A. Hajimiri, T. Lee, "A General Theory of Phase Noise in Oscillators," *IEEE Journal of Solid-State Circuits* (in press).