Linearity Analysis of CMOS for RF Application

Sanghoon Kang, Byounggi Choi, and Bumman Kim, *Senior Member, IEEE*

***Abstract*—The linearity of CMOS has been analyzed using the Taylor series. Transconductance and output conductance are two dominant nonlinear sources of CMOS. At a low frequency, the transconductance is a dominant nonlinear source for a low load impedance, but for a usual operation level impedance the output conductance is a dominant nonlinear source. Capacitances and the substrate network do not generate any significant nonlinearity, but they suppress output-conductance nonlinearity at a high frequency because output impedance is reduced by the capacitive shunts, and output voltage swing is also reduced. Therefore, above 2–3 GHz, the transconductance becomes a dominant nonlinear source for a usual load impedance. If these capacitive elements are tuned out for a power match, the behavior becomes similar to the low-frequency case. As gate length is reduced, the transconductance becomes more linear, but the output conductance becomes more nonlinear. At a low frequency, CMOS linearity is degraded as the gate length becomes shorter, but at a higher frequency (above 2–3 GHz), linearity can be improved.**

***Index Terms*—CMOS, linearity.**

I. Introduction

THERE IS A strong interest in using CMOS technology for RF and microwave circuits. To support the growing needs of RF circuits, the various characteristics of CMOS have been evaluated using various device models. However, they are usually small-signal behaviors, such as small-signal gain, noise figure, etc. More recently, the large signal characteristics of CMOS, such as power gain, power generation, efficiency, and mixer operation have been intensively studied. However, linearity is always one of the most important issues for RF circuit design.

Some research considers only the transconductance in CMOS linearity analysis, while other effects are neglected [1]. For a short gate CMOS, however, the output conductance becomes an important nonlinear source and must be included in CMOS analysis. Capacitances and substrate effects are also far from negligible at high frequency.

We have studied the nonlinear characteristics of CMOS using Taylor-series analysis and a BSIM3–based model [2]. We have analyzed the output-conductance nonlinearity of CMOS and have compared it with the transconductance nonlinearity. The linearity trend with a down-scaled device is also described in Section II. The capacitance contribution for linearity and substrate network behavior are also analyzed, and their effects are discussed in Section III.

Manuscript received May 13, 2002; revised September 18, 2002. This work was supported in part under the Brain Korea 21 Project, by the Agency for Defence Development, and by Hynix.

The authors are with the Department of Electronic and Electrical Engineering, Microwave Application Research Center, Pohang University of Science and Technology Kyoungbuk 790-784, Korea. (e-mail: kangnaru@postech.ac.kr; bmkim@postech.ac.kr)

Digital Object Identifier 10.1109/TMTT.2003.808709

II. Transconductance and Output-Conductance Nonlinearities

The BSIM3 model is widely used in circuit simulation tools, but it requires a large data set for accurate modeling and is not suitable for a simple analysis. However, the model contains all the information for nonlinear behavior. Taylor series are usually used for weakly nonlinear circuit analysis because it is simple. However, it cannot be applicable to highly nonlinear applications such as power amplifiers. In this paper, Taylor-series analysis is employed and the expansion coefficients are extracted from the BSIM3 model. The drain current in Taylor expansions can be expressed as follows:

$$\begin{aligned} i_{\rm ds}(v_{\rm GS}, v_{\rm DS}) =& I_{\rm DS}(V_{\rm GS}, V_{\rm DS}) + G_m v_{\rm gs} + G_d v_{\rm ds} \\ &+ G_{m2} v_{\rm gs}^2 + G_{md} v_{\rm gs} v_{\rm ds} + G_{d2} v_{\rm ds}^2 \\ &+ G_{m3} v_{\rm gs}^3 + G_{m2d} v_{\rm gs}^2 v_{\rm ds} \\ &+ G_{md2} v_{\rm gs} v_{\rm ds}^2 + G_{d3} v_{\rm ds}^3 + \cdots. \end{aligned} \tag{1}$$

Assuming that the drain is shorted at a signal frequency, all output-conductance terms (G_d, G_{d2}, G_{d3}, ...) and cross modulation terms (G_{md}, G_{m2d}, G_{md2}, ...) are vanished and only transconductance terms (G_m, G_{m2}, G_{m3}, ...) remain. In this case, the third-order intercept point (IP3) of gate voltage amplitude given as follows in (2) has been used as a device linearity criterion in many previous reports [1]:

$$\mathrm{IP3} = \sqrt{\frac{4}{3}\frac{G_m}{G_{m3}}}. \tag{2}$$

This model is not accurate enough, however, because it does not consider output-conductance nonlinearity. To compare output-conductance nonlinearity with transconductance nonlinearity, the output nonlinear currents from the two components are calculated. For typical CMOS processes, the higher order terms and cross-modulation terms in (1) are very small and can be ignored to simplify the calculation. In this approximation, the third-order intermodulation currents caused by the transconductance and output-conductance nonlinearities, at a low frequency, are given by

$$i_{\rm IM3trans} = \frac{3}{4} G_{m3} A^3 \frac{G_{\rm load}}{G_d + G_{\rm load}} \tag{3}$$

$$i_{\rm IM3cond} = \frac{3}{4} G_{d3} v_{\rm ds}^3 \frac{G_{\rm load}}{G_d + G_{\rm load}} \tag{4}$$

where A is the fundamental voltage amplitude at the gate and $v_{\rm ds}$ is the fundamental voltage at the drain, which can be expressed as

$$v_{\rm ds} = \frac{G_m A}{G_{\rm load} + G_d}. \tag{5}$$

0018-9480/03$17.00 © 2003 IEEE

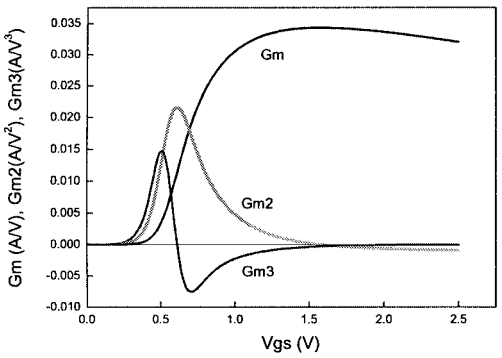

Fig. 1. TSMC 0.25-μm NMOS transconductance.

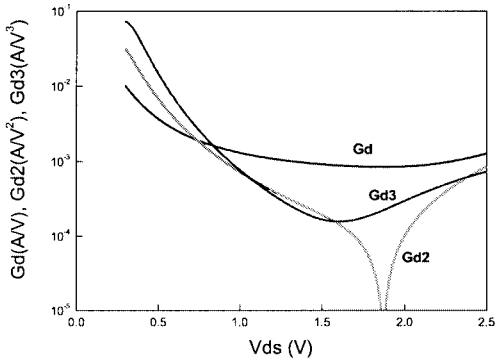

Fig. 2. TSMC 0.25-μm NMOS output conductance.

The TSMC Company Ltd., Taiwan, R.O.C., 0.25 μm$\times$ 10 μm$\times$ 10 NMOS BSIM3v3 model is used for this study [4]. All Taylor-series parameters are extracted from the dc *IV* curve of the model. The extracted parameters are shown in Figs. 1 and 2. Fig. 1 shows G_m, G_{m2}, and G_{m3} versus $V_{\rm gs}$ for a constant $V_{\rm ds}$ ($=$ 1.5 V). We selected the gate bias $V_{\rm gs} = 0.8$ V for comparison because this point is where the transconductance nonlinearity is large with sufficient gain. Fig. 2 shows G_d, G_{d2}, and G_{d3} at the selected gate bias voltage. $V_{\rm ds} = 1.5$ V is selected as a bias point because this is usual bias voltage for 0.25 μm NMOS considering load resistance voltage drop and the output-conductance nonlinearity of the MOS is near minimum.

Using these parameters and (3)–(5), we calculate the output third-order intermodulation current of the circuit given in Fig. 3 for various loads and compare them with the results of harmonic-balance simulation. Two-tone source voltage of 2 mV ($=-50$ dBm in 50 Ω) is injected to ensure a weakly nonlinear operation. The result is shown in Fig. 4, where the intermodulation currents are shown in a 20log scale. Only with G_m, G_{m3}, G_d, and G_{d3} coefficients, the calculation result matches the simulation result very well. When we include G_{md} and other higher order components, the calculations predict the simulated results even better. However, their effects are small and can be neglected in the saturation region of the MOSFET. However in the linear region, the cross-modulation terms are large and should be included to achieve an accurate analysis.

The nonlinear currents are dependent on load impedance. When drain voltage swing is small at a low load resistance, the transconductance is the dominant nonlinear source. At a high load resistance, however, the output conductance is the dominant nonlinear source. The decreases of IM3 at a high load resistance are shunting effect of G_d and $G_{\rm load}$. The output third-order intercept point (OIP3) and power gain for a 300-Ω source resistance are calculated. The results are depicted in Fig. 5. The maximum OIP3 is obtained at a load resistance lower than the value for the maximum gain ($= R_{\rm gain}$). The best linearity is achieved at the load resistance where the output-conductance nonlinearity increases to a level comparable to the transconductance nonlinearity ($= R_{\rm lin}$). Up to $R_{\rm lin}$, the fundamental and intermodulation currents ($= i_{\rm IM3}$) remain the same and their powers increase at the same rate, proportional to the load resistance and OIP3 increases. The fundamental current stays constant up to $R_{\rm gain}$, but the intermodulation current from the output-conductance nonlinearity ($= i_{\rm IM3cond}$) increases

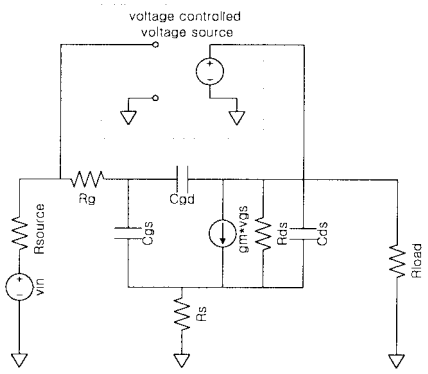

Fig. 3. NMOS model for linearity analysis.

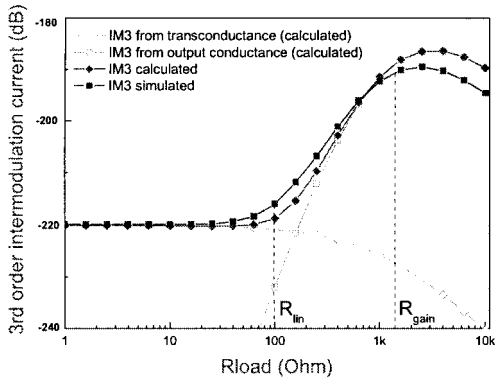

Fig. 4. Output IM3 currents versus load resistance.

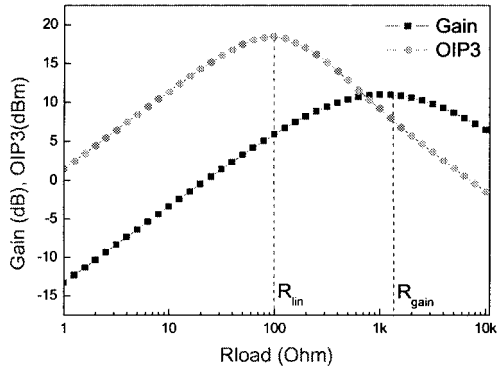

Fig. 5. Gain and OIP3 versus load resistance.

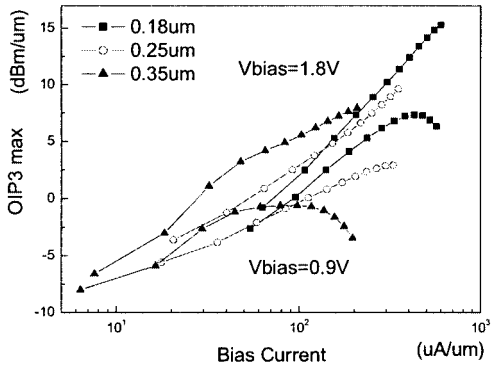

Fig. 6. Linearity of various gate-length MOS (TSMC process).

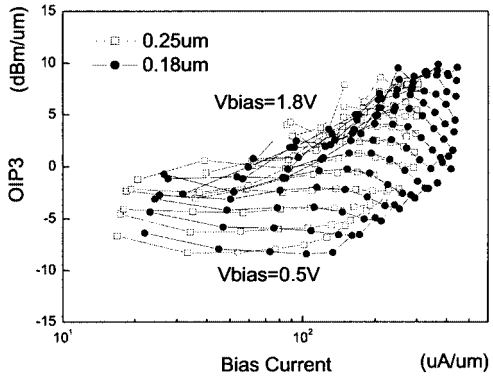

Fig. 7. Linearity of various gate-length MOS (Hynix process).

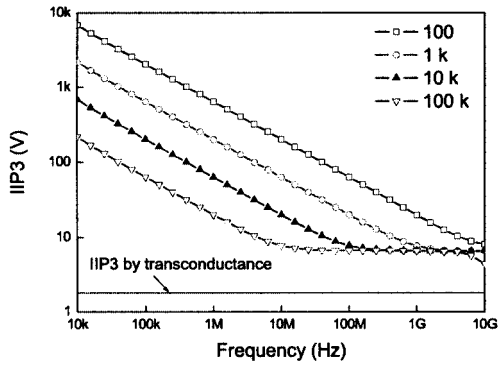

Fig. 8. IIP3 from $C_{\rm gs}$ and $C_{\rm gd}$ nonlinearity for various $R_{\rm source}$.

rapidly in this region, and OIP3 decreases. Above $R_{\rm gain}$, $i_{\rm IM3}$ and power gain both decrease. In many RF applications, the load impedance is generally matched to the device output impedance for a maximum power transfer. If we desire a highly linear operation, however, the output impedance should be reduced from the power-matching impedance. It is desirable to select an output load between $R_{\rm lin}$ and $R_{\rm gain}$ for good RF performances.

To understand the process-dependent linearity trend, we evaluate the maximum OIP3 for TSMC 0.35-, 0.25-, and 0.18-μm processes. The total gatewidth is 100 μm with ten fingers. Since the linearity varies with output load resistance, we calculate the maximum OIP3s and corresponding load resistances for various current levels and drain bias voltages. As the gate length shortens, the transconductance becomes more linear, but the output conductance becomes more nonlinear. As a result, the maximum OIP3 of CMOS does not change for the gate-length variations, as shown in Fig. 6. A shorter gate MOSFET delivers a little better linearity at a high current density, but the improvement is negligible and at a lower current density, the same or a slightly worsened linearity is achieved. However, the maximum OIP3 of a short MOSFET occurs at a lower load impedance, and for usual operation level load impedance, the OIP3 is degraded as the gate length is reduced. If the input third-order intercept point (IIP3) is considered, a shorter gate device is much more nonlinear at the same bias current level because of its higher gain.

To verify the above conclusions from measured data, we evaluate 0.18- and 0.25-μm NMOS devices from Hynix. We measure $v_{\rm gs} - i_{\rm ds}$ and $v_{\rm ds} - i_{\rm ds}$ curves using an Agilent 4155 parameter analyzer, and extract G_m, G_{m3}, G_d, and G_{d3} from the measured dc *IV* data. To obtain smooth data from the measurements, we use a data-processing program based on MATLAB. Maximum OIP3s are calculated for various current densities and bias voltages using (3) and (4), and are depicted in Fig. 7. As shown, the OIP3 of the MOSFET depends only on the current density and bias voltage, but does not depend on the gate length. These trends are identical to the simulation results.

III. High-Frequency Effects due to Capacitive Components

Previous experimental report shows that the input impedance of CMOS is a major source of nonlinearity at a high frequency [3]. The input impedance consists of two nonlinear components, gate–source capacitance ($C_{\rm gs}$) and feedback component through gate–drain capacitance ($C_{\rm gd}$). The effects of $C_{\rm gs}$ and $C_{\rm gd}$ on the input nonlinearities are evaluated numerically for various source resistances. To eliminate the $C_{\rm gd}$ Miller effect, the drain is shorted. We simulated IIP3 voltage from the gate nonlinearity and the results are shown in Fig. 8. The IIP3 voltage from the $C_{\rm gs}$ and $C_{\rm gd}$ nonlinearities are significantly larger than that from transconductance nonlinearity for all source resistances and frequencies, i.e., the intermodulation components generated by $C_{\rm gs}$ and $C_{\rm gd}$ are much smaller than that caused from transconductance.

To eliminate the $C_{\rm gd}$ effect, we add a unity gain voltage-controlled voltage source, with input port to the gate and output port to the drain, as shown in Fig. 3. The IIP3 voltage is reduced by only approximately 1 dB, indicating that the effect of the $C_{\rm gd}$ nonlinearity is smaller than that of $C_{\rm gs}$. However, the intermodulation components at the gate are increased for a large load resistance because $C_{\rm gd}$ forms a feedback loop for the harmonics, which are generated at the drain, and disturbs device linearity. When we simulate the input nonlinearity with $C_{\rm gd}$ zero, the gate voltage intermodulation components are reduced by approximately 8 dB at minimum. For the other capacitance, drain–source capacitance ($C_{\rm ds}$) is small and remains almost constant at the saturation region, and its effect is negligible. In summary, the capacitive elements do not generate any significant harmonics, but $C_{\rm gd}$ influences the harmonics by the feedback.

Also, $C_{\rm gd}$ and $C_{\rm ds}$ can reduce the output impedance at a high frequency and effectively reduce the output-conductance non-

linearity due to the reduced voltage swing. The output IM3 current modified by these capacitances can be obtained using the following equations:

$$v_d = \frac{-G_m}{j\omega C_{\rm gd}(1-\frac{1}{A_v}) + j\omega C_{\rm ds} + G_d + G_{\rm load}} \cdot v_g \quad (6)$$

$$v_g = \frac{1}{1 + Z_s\left(j\omega C_{\rm gs} + j\omega C_{\rm gd}(1-A_v)\right)} \cdot v_{in} \quad (7)$$

$$A_v = \frac{-G_m + j\omega C_{\rm gd}}{G_{\rm load} + G_d + j\omega C_{\rm ds} + j\omega C_{\rm gd}} \quad (8)$$

where A_v is the voltage gain from the gate to drain. The intermodulation currents are given by

$$\begin{aligned} |i_{\rm IM3trans}| &= \left|\frac{3}{4}\frac{G_{\rm load}}{G_{\rm load}+Y_{\rm out}} G_{m3} v_g^3\right| \\ \angle i_{\rm IM3trans} &= \angle\left[\frac{G_{\rm load}}{G_{\rm load}+Y_{\rm out}} v_g\right] \end{aligned} \quad (9)$$

$$\begin{aligned} |i_{\rm IM3cond}| &= \left|\frac{3}{4}\frac{G_{\rm load}}{G_{\rm load}+Y_{\rm out}} G_{d3}\cdot v_d^3\right| \\ \angle i_{\rm IM3cond} &= \angle\left[\frac{G_{\rm load}}{G_{\rm load}+Y_{\rm out}} G_{d3} v_d\right] \end{aligned} \quad (10)$$

$$Y_{\rm out} = j\omega C_{\rm gd}(1-\frac{1}{A_v}) + j\omega C_{\rm ds} + G_d. \quad (11)$$

The drain intermodulation voltage is feedback to the gate through $C_{\rm gd}$ and the feedback effect can be calculated as

$$i_{\rm IM3transmod} = \frac{i_{\rm IM3trans}}{1 - A_v\beta} \quad (12)$$

$$i_{\rm IM3condmod} = \frac{i_{\rm IM3cond}}{1 - A_v\beta} \quad (13)$$

where β is the feedback factor given by

$$\beta = \frac{j\omega C_{\rm gd}}{j\omega(C_{\rm gs}+C_{\rm gd}) + \dfrac{1}{R_s + R_g}}. \quad (14)$$

The IM3 output current is calculated up to 6 GHz using these simple equations and is compared with the simulation results for various load resistances. They agree very well. The calculated result shows that G_{m3} and G_{d3} are the dominant nonlinear source up to the frequency. At a low frequency, G_{d3} is the dominant device nonlinear source for a usual load impedance, but at a high frequency (above 4 GHz), the transconductance becomes more important for all load impedance.

The effects of conductive substrate and drain–source–substrate junction diode capacitance on CMOS linearity have been studied using the model proposed in [4] and depicted in Fig. 9. In the substrate network, the resistances are considered as linear elements, but the source–substrate and drain–substrate diodes are always reverse biased, working as nonlinear capacitors. The p-n junction capacitors can be approximated by (15) as follows:

$$C_j = \frac{C_{j0}}{\left(1 + \dfrac{v_R}{V_j}\right)^{m_j}} \quad (15)$$

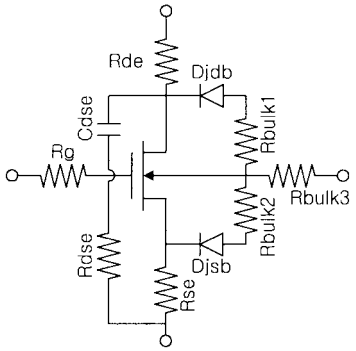

Fig. 9. MOSFET substrate network model.

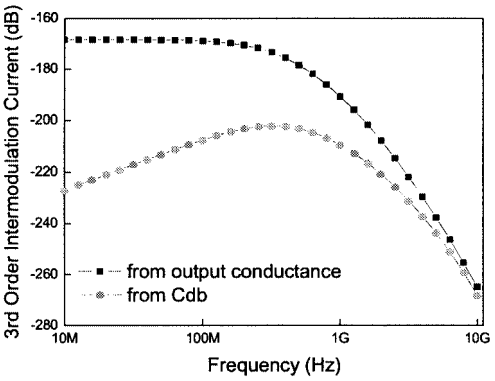

Fig. 10. Third-order intermodulation current from output conductance and C_{db}.

in which C_{j0} is the junction capacitance at zero bias. The capacitance nonlinear current can then be calculated as follows [5]:

$$\begin{aligned} i_{\rm out} = C_j \cdot \frac{dv_r}{d_t} - C_j \cdot \frac{1}{2\left(1+\dfrac{V_R}{V_j}\right)} \cdot \frac{m_j}{V_j}\frac{d}{dt}\left(v_r^2\right) \\ + C_j \cdot \frac{1}{6\left(1+\dfrac{V_R}{V_j}\right)^2} \cdot \frac{m_j(m_j+1)}{V_j^2}\frac{d}{dt}\left(v_r^3\right) + \cdots. \end{aligned} \quad (16)$$

Using this equation and parameters from [4], we calculate the output third-order intermodulation current generated from the drain–substrate junction capacitor ($= i_{\rm IM3}{}_{cdb}$). In Fig. 10, we compare it with the intermodulation current generated from output conductance ($= i_{\rm IM3condmod}$). $i_{\rm IM3}{}_{cdb}$ is much less than $i_{\rm IM3condmod}$ and can be neglected with a little error up to a moderate frequency. However, the difference is reduced proportionally at a higher frequency and, above 10 GHz, it may be necessary to include the nonlinearity for an accurate analysis. However, this frequency can be varied according to the pad size and substrate doping level.

For a simple calculation, we assume that C_{db} is constant and modify (6)–(13) to include the following substrate admittance:

$$Y_{\rm sub} = \frac{1}{\dfrac{1}{R_{\rm bulk3} + \dfrac{j\omega C_{sb}}{1 + j\omega C_{sb} R_{\rm bulk2}}} + R_{\rm bulk1} + \dfrac{1}{j\omega C_{db}}}. \quad (17)$$

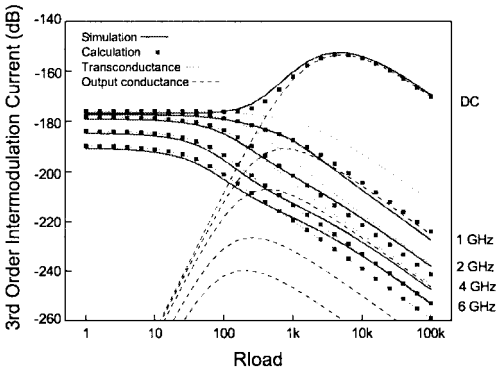

Fig. 11. Third-order intermodulation current for CMOS with all capacitive elements.

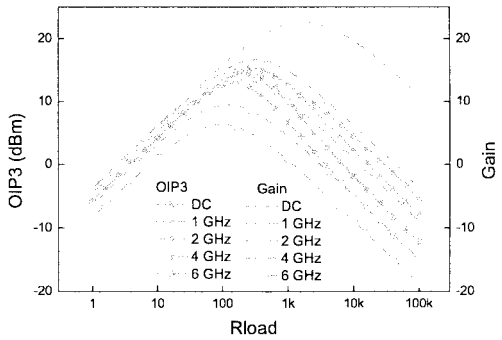

Fig. 12. Gain and OIP3 for CMOS including all capacitive elements.

The calculated results using our model are compared with harmonic-balance simulation results. As shown in Fig. 11, there are good agreements between the two results. The load-dependent nonlinear currents from the transconductance and output conductance, shown in Fig. 11, include all the capacitive effects and substrate network. The OIP3 and gain curves are depicted in Fig. 12. Due to the capacitive shunting effect, the output voltage swing is reduced and the nonlinear current of the output conductance is reduced. Therefore, at a high frequency, $R_{\rm gain}$ is reduced, but $R_{\rm lin}$ remains almost constant. When the substrate network is included, the crossover frequency where the transconductance nonlinearity of the 0.25-μm MOSFET becomes larger than the output-conductance nonlinearity is further reduced to approximately 2 GHz, and OIP3 is also significantly reduced. Therefore, at a high frequency, the optimum load impedance for best linearity can be higher than that for power matching.

In many RF applications, the input and output ports are matched for maximum power transfer. In this case, the capacitive elements are tuned out and the linearity characteristics of NMOS can differ from the previous results for resistive loads. In Fig. 13, we show the IM3 output current when the output capacitances are tuned out by a parallel inductor at each frequency. Voltage gain is increased and the crossover frequency is also increased. The difference is clearly seen at 2 GHz compared with Fig. 11. The output-conductance nonlinearity is the dominant nonlinearity source for the tuned matching load.

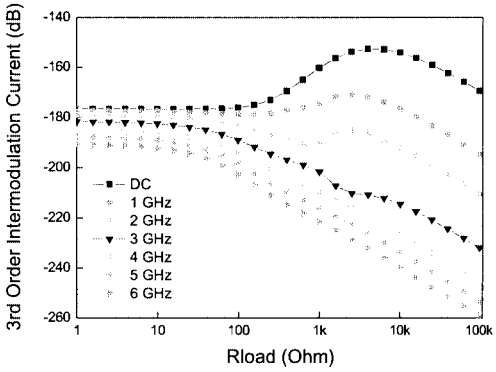

Fig. 13. Third-order intermodulation current when output capacitance is tuned out.

IV. Conclusions

CMOS linearity has been analyzed and we have found that G_{m3} and G_{d3} are the dominant nonlinear sources for the frequency band we have studied (up to 6 GHz). Also, their behaviors are described using simple equations we have derived. Internal capacitances and substrate network are also included in the calculation. The capacitive components are not the sources of nonlinearity, but they reduce the high-frequency gain and output voltage swing. Therefore, at a high frequency, the transconductance nonlinearity becomes a dominant source. However, at a low frequency, or for a tuned matching circuit where the capacitive shunting effects are removed, the output conductance is the dominant nonlinear source. The crossover frequency, between the low- and high-frequency behaviors, for a 0.25-μm MOSFET is around 2 GHz. As the gate length is reduced, the transconductance becomes more linear, but the output-conductance nonlinearity is enhanced. Further, CMOS with short gate length is more nonlinear at a low frequency, but at the frequency region where the capacitive element shunting effect becomes important, CMOS becomes more linear as the gate length is reduced. These descriptions for the IMD3 generation mechanism of CMOS will be a useful guideline for RF circuit design.

Acknowledgment

The authors would like to thank Hynix, Ichon, Korea, for supplying the device samples.

References

[1] P. H. Woerlee *et al.*, "RF-CMOS performance trends," *IEEE Trans. Electron Devices*, vol. 48, pp. 1776–1782, Aug. 2001.
[2] S. Kang, B. Choi, and B. Kim, "Linearity analysis of CMOS for RF application," in *IEEE RFIC Symp. Dig.*, 2002, pp. 363–366.
[3] C. Wang *et al.*, "A nonlinear capacitance cancellation technique and its application to a CMOS class AB power amplifier," in *IEEE RFIC Symp. Dig.*, 2001, pp. 39–42.
[4] "TSMC 0.25 μm Mixed-signal salicide RF Spice models," TSMC Company Ltd., Taiwan, R.O.C., Sept. 2000.
[5] W. Piet and S. Willy, *Distortion Analysis of Analog Integrated Circuits*. Norwell, MA: Kluwer, 1998.

Sanghoon Kang was born in Pusan, Korea, in 1972. He received the B.S. and M.S. degrees in electrical engineering in 1996 and 1998, respectively, from Pohang University of Science and Technology, Pohang, Korea, where he is currently working toward the Ph.D. degree.

His research interests include CMOS RFIC design, in particular, mixer linearity and noise.

Byounggi Choi was born in Daegu, Korea, in 1975. He received the B.S. degree in electronics engineering from Sunggyunkwan University, Seoul, Korea, in 2000, and is currently working toward the M.S. degree at the Pohang University of Science and Technology, Kyoungbuk, Korea.

His current research interest is the linearity of mixers.

Bumman Kim (S'77–M'78–SM'97) received the Ph.D. degree in electrical engineering from Carnegie-Mellon University, Pittsburgh, PA, in 1979.

From 1978 to 1981, he was engaged in fiber-optic network component research at GTE Laboratories Inc. In 1981, he joined the Central Research Laboratories, Texas Instruments Incorporated, where he was involved in development of GaAs power FETs and monolithic microwave integrated circuits (MMICs). He has developed a large-signal model of a power FET, dual-gate FETs for gain control, high-power distributed amplifiers, and various millimeter-wave MMICs. In 1989, he joined Pohang University of Science and Technology, Pohang, Korea, where he is a Professor in the Electronic and Electrical Engineering Department, and Director of the Microwave Application Research Center, where he works on device and circuit technology for MMICs. Currently, he is a Visiting Professor of Electrical Engineering at California Institute of Technology, Pasadena. He has authored more than 100 published technical papers.

Dr. Kim is a member of the Korean Academy of Science and Technology and Academy of Engineering of Korea.