THE PERFORMANCE LIMITING FACTORS AS RF MOSFETs SCALING DOWN

Y. H. Wu, Albert Chin, C. S. Liang, and C. C. Wu
Dept. of Electronics Eng., National Chiao Tung Univ., Hsinchu, Taiwan

ABSTRACT

The measured RF performance of 0.5, 0.25, and 0.18 μm MOSFETs gradually saturates as scaling down, which can be explained by the derived analytical equation and simulation. The overlap C_{gd} and non-quasi-static effect are the main factors but scales much slower than L_g.

INTRODUCTION

Although Si RF MOSFETs has the advantages of rapid technology evolution and low production cost, it is still not clear where is the limitation of MOSFETs [1], and whether Si BJT [2] or even III-V technology should be used at higher frequencies. In this paper, we have analyzed the fabricated 0.5, 0.25, and 0.18 μm MOSFETs, and discussed performance limiting factors as scaling down using our derived analytical equation and numerical device simulation. We have found that the RF performance improvement gradually saturates as scaling down, which is observed by both experimental data and our analysis. The source-drain overlap capacitance (C_{gd}) is the key factor for G_{max} and f_{max}; unfortunately, it is difficult to proportionally scale down as L_g due to lateral diffusion of source-drain implants. The non-quasi-static (NQS) effect will also reduce the H_{21}, f_t, and maximum available gain (MAG). Our work can help to understand the performance limitation of MOSFETs scaling and further choose of device operated at high frequencies.

EXPERIMENTAL

Multiple fingered 0.5, 0.25, and 0.18 μm MOSFETs are fabricated on standard ~10 Ω-cm Si substrate with gate width of 200-250 μm and on-wafer probe layout. The multiple gate fingers with low resistivity $CoSi_2$ [3] can achieve a reasonable power level and reduce the extrinsic gate resistance that is important for G_{max} and f_{max}. Then, S-parameters were measured up to 18 GHz using a CASCADE on-wafer probe, a network analyzer, and de-embedded from dummy devices. A matrix of different size of transistors and capacitors is used to extract device parameters for further analysis using modified BSIM3v3 MOSFETs model in SPICE.

RESULTS AND DISCUSSION

The measured frequency response of H_{21} and G_{max} is plotted in Fig. 1 and summarized in Table 1. It is important to notice that the measured H_{21}, f_t, G_{max}, and f_{max} gradually saturate as device scaling down. The saturation rate is faster for G_{max} and a reducing f_{max} is even observed.

Furthermore, the measured H_{21} and f_t are about 50% lower than the calculated value from conventional equation of $g_m/2\pi C_{gs}$ or $v_{sat}/2\pi(L_g\text{-}2L_{ov})$, where L_{ov} is the gate-drain overlap length. We have therefore derived a more accurate H_{21} and f_t (at H_{21}=1) equations using modified BSIM3v3 equivalent circuit model and including the NQS effect.

0-7803-6280-2/00/$10.00 © 2000 IEEE

Table I. Measured and calculated RF data.

measured/ calculated values	mea. H_{21} (dB) 4GHz	mea. f_T (GHz) H_{21} =1	cal. f_T (GHz) H_{21} =1	mea. f_{max} (GHz) MAG =1	mea. f_{max} (GHz) MSG =1	Cal. f_{max} (GHz) MSG =1	mea. G_{max} (dB) 4GHz	cal. G_{max} (dB) 4GHz
0.5-μm	14.7	25	23	20	82	80	13.0	13.9
0.25-μm	19.7	42	38	18	119	127	15.0	15.9
0.18-μm	22.2	58	56	17	161	171	16.3	18.0

$$H_{21} = \sqrt{[\frac{g_m}{w(C_{gs}+C_{gd})}]^2 + (g_m R_{nqs} - \frac{C_{gd}}{C_{gs}+C_{gd}})^2} \tag{1}$$

$$f_t = \frac{g_m}{2\pi(C_{gs}+C_{gd})} \frac{1}{\sqrt{1-\left(g_m R_{nqs} - \frac{C_{gd}}{C_{gs}+C_{gd}}\right)^2}} = \frac{v_{sat}}{2\pi(L_g+L_{ov})} \frac{1}{\sqrt{1-\left(g_m R_{nqs} - \frac{L_{ov}}{L_s+L_{ov}}\right)^2}} \tag{2}$$

$$G_{\max} = \left|\frac{S_{21}}{S_{12}}(K-\sqrt{K^2-1})\right| = \left|\frac{jwC_{gd} - g_m[1+jwR_{nqs}(C_{gd}+C_{gs})]}{jwC_{gd}}\right| \left|(K-\sqrt{K^2-1})\right| \tag{3}$$

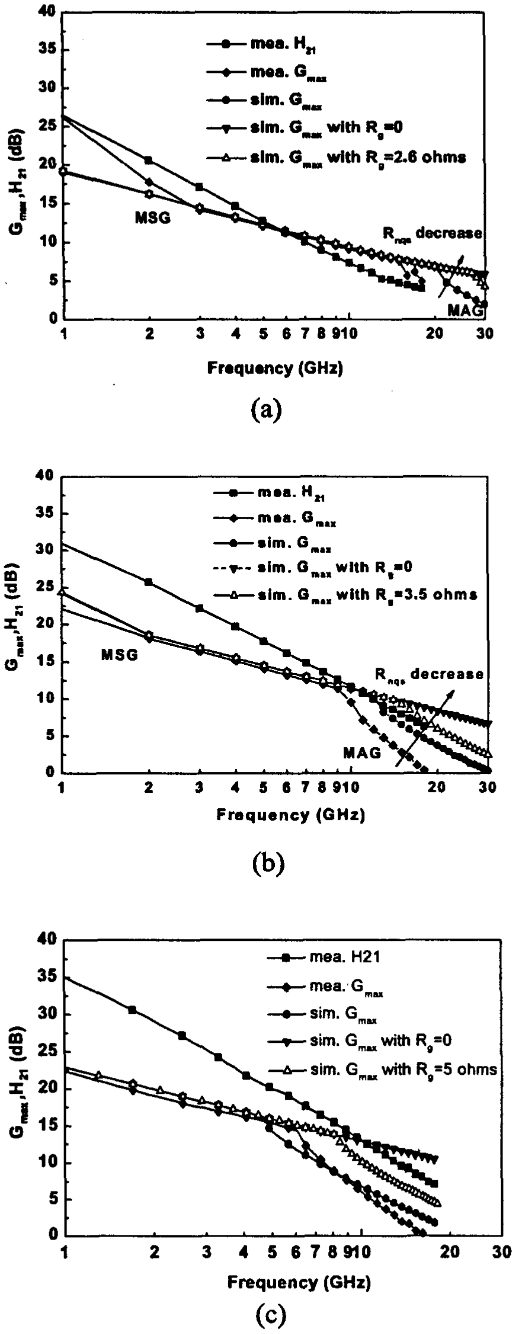
$$f_{\max,MSG=1} = \frac{1}{2\pi} \frac{g_m}{C_{gd}} \frac{1}{\sqrt{2g_m R_g C - (g_m R_g C)^2}} \qquad C = 1 + \frac{C_{gs}}{C_{gd}} \tag{4}$$

Although the R_{nqs} related term in H_{21} is negligible at low frequency, it becomes more important as increasing frequency near f_t. Good matching between measured and simulated f_t in Table 1 can only be obtained by considering the NQS effect. Because of the additional term, f_t increases slower than 1/ L_g scaling down.

Similar large difference of 300%-350% exists in the measured and calculated f_{max} using the well-known equation of $(f_t/8\pi R_g C_{gd})^{1/2}$. This difference is because the above equation is derived from the unilateral gain with a constant gain roll-off while G_{max} changes to 30-40dB/decade decrease in MAG.

To further analyze the frequency response, we have also derived G_{max} and f_{max} by using the equivalent circuit modeling and including the NQS effect. From derived G_{max}, C_{gd} related pole gives the 10dB/decade G_{max} roll-off in MSG, while the large slope of ~30-40dB/decade in MAG is due to additional poles in K or the NQS effect on g_m.

Although similar method can be used to calculate f_{max} at G_{max}=1, unfortunately, no analytical solution can be derived for f_{max}. In contrast, analytical f_{max} at MSG=1 can be obtained when $\left|(K-\sqrt{K^2-1})\right|=1$, we have therefore analyzed $f_{max,MSG=1}$ to obtain a better understanding of device design parameters on $f_{max,MSG=1}$. Good agreement between the measured and calculated $f_{max,MSG=1}$ is achieved and shown in Table 1. The primary parameter for $f_{max,MSG=1}$ increase is due to the g_m increase and C_{gd} decrease. In fact, C_{gd} is dominated by the C_{gd} that is difficult to proportionally scale down with L_g.



(a)

(b)

(c)

Fig.1 Gain-frequency response for measured and simulated (a) 0.5, (b) 0.25, and (c) 0.18 μm MOSFETs.

We have also used numerical device simulation for further analysis. We have studied the NQS effect on G_{max} and f_{max}. As shown in Fig. 1, the MAG increases with decreasing R_{nqs} and eventually gives G_{max} the same 10dB/decade roll-off as MSG when R_{nqs} equals 0. Therefore, the NQS effect is responsible for the transition from MSG to MAG. Because R_{nqs} is inversely related to C_{gs}, a higher dielectric or thinner gate thickness is required to improve the high frequency gain.

On the other hand, G_{max} has a simple analytical solution in the most useful MSG region for amplifier design. Because the $R_{nqs}(C_{gs}+C_{gd})$ related zeros are effective only at high frequencies, G_{max} in MSG can be further simplified and expressed by $g_m/\omega C_{gd}$ or $v_{sat}/\omega L_{ov}$. The numerical simulation result is shown in Fig. 2. It is clear that the reduction of C_{gd} leads to a higher G_{max} and f_{max}. However, the difference between the ideal $2C_{ox}Wt_{ox}$ and the measured data is larger as scaling down.

Here, a minimum C_{gd} of $C_{ox}WL_{ov}$ ($L_{ov}=2t_{ox}$) [4] is required in order to develop a reproducible and manufacturable process, where C_{ox} and t_{ox} are the gate capacitance and oxide thickness, respectively. Although down scaling gives a smaller L_g and a higher C_{ox}, limited G_{max} improvement in MSG is due to the slower scalable L_{ov}. The reason for L_{ov} failing to follow t_{ox} scaling down in deep sub-μm devices is due to the lateral diffusion from source and drain impurities. High temperature annealing after source and drain implantation is necessary to reduce the junction leakage but largely increases the lateral diffusion. The formation of silicide junction also requires high temperature RTA. Because of the combined small G_{max} and K factor improvement, limited f_{max} improvement as device scaling down can be expected.

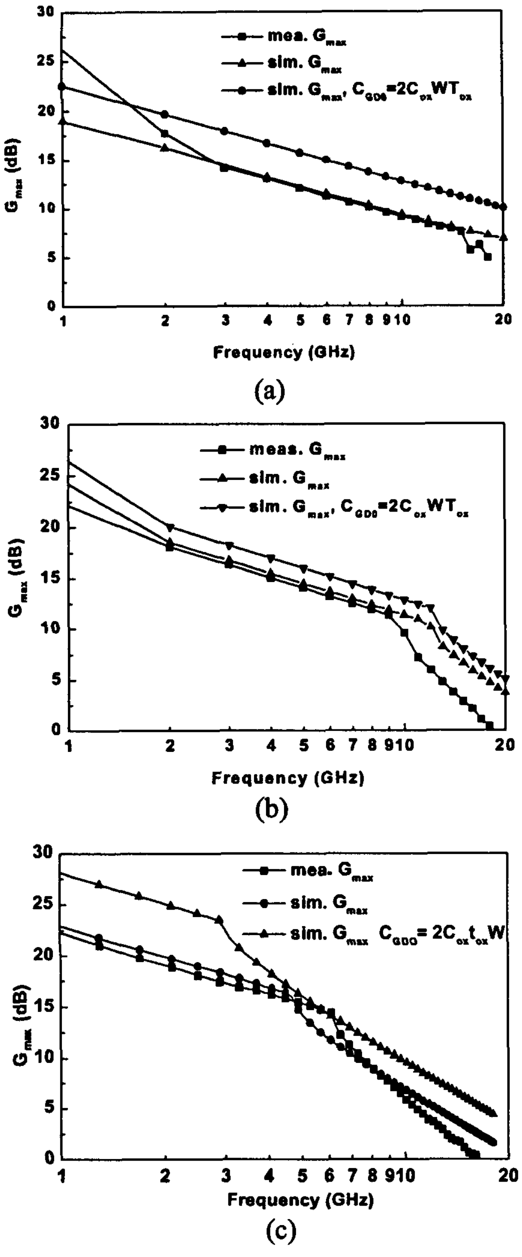


Fig.2 The effect of reducing C_{gd} on gain-frequency response for (a) 0.5, (b) 0.25, and (c) 0.18 μm MOSFETs.

The smaller increase of measured G_{max} than calculated value in Table 1 as down scaling may be due to the parasitic effect neglected in our device model.

CONCLUSIONS

We have analyzed the RF performance of 0.5, 0.25, and 0.18 μm RF MOSFETs. Because of the NQS effect, H_{21}, f_t, and MAG improve slower than L_g decreasing. The small $G_{max,at\ 4GHz}$ and f_{max} improvement as scaling down is primary due to the feedback path of C_{gd}.

REFERENCES

[1] M. C. Ho, F. Brauchler, and J. Y. Yang, "Scalable RF Si MOSFET distributed lumped element model based on BSIM3v3," Electronics Lett., vol. 33, no. 23, pp. 1992-1993, 1997.

[2] S. Niel, O. Rozeau, L. Ailloud, C. Hernandez, P. Llinares, M. Guillermet, J. Kirtsch , A. Monroy, J. de Pontcharra, G. Auvert, B. Blanchard, M. Mouis, G. Vincent, and A. Chantre, "A 54 GHz f_{max} implanted base 0.35μm single-polysilicon bipolar technology," in *IEDM Tech. Dig.*, 1997, pp. 807-810.

[3] Y. H. Wu, W. J. Chen, S. L. Chang, A. Chin, S. Gwo, and C. Tsai, "Improved electrical characteristics of $CoSi_2$ using HF-vapor pretreatment," *IEEE Electron Device Lett.*, vol. 20, no. 5, 200, 1999.

[4] T. Y. Chan, A. T. Wu, P. K. Ko, and C. Hu, "Effects of the gate-to-drain/source overlap on MOSFET characteristics," *IEEE Electron Device Lett.*, vol. 8, pp. 326-328, 1987.