publisher prior to publication. The final version is available at http://dx.doi.org/10.1109/LATW.2017.7906750

Analysis of the Implications of Stacked Devices in Nano-Scale Technologies for Analog Applications

Ismael Lomelí-Illescas, Sergio A. Solis-Bustos
Intel Tecnología de México S.A. de C.V.
Zapopan, Jalisco, 45019 México
e-mail: ismael.lomeli@intel.com

José E. Rayas-Sánchez
Department of Electronics, Systems, and Informatics,
ITESO - The Jesuit University of Guadalajara, Tlaquepaque,
Jalisco, 45604 Mexico. http://desi.iteso.mx/erayas/

***Abstract* — In this work, a methodology to assess the implications on the performance of analog circuits due to the use of stacked devices in current nano-scale technologies is presented. To evaluate the usage of stacked devices, the characteristic curves of transistors implemented with a different amount of transistors in stack are obtained and compared to those of a single device. The effects of using stacked devices are further studied with the implementation of a current mirror and the implementation of two different layout topologies, discussing their tradeoffs, advantages and drawbacks. Our methodology facilitates designers to develop a good understanding of the characteristics and limitations of a particular physical design before silicon is back for laboratory testing.**

***Index Terms* — Analog layout, channel modulation, interdigitated layout, leakage, stacked devices, stack effect.**

I. Introduction

In current nano-scale technology processes, one of the main challenges in the area of analog circuit design is the implementation of high performance circuits using devices for digital applications. Reduction of device dimensions and power supply voltages, as well as the limitations to define the transistor widths (now restricted to discrete values), are forcing designers to implement new structures to emulate the correct analog behavior [1]. One of the most commonly used solutions, but that not deeply studied in nano-scale technology processes, consists of using transistors placed in stack. In this paper, a methodology to assess the implication on the use of stacked devices is presented. The analysis of the effects of using stacked arrays of transistors instead of standalone transistors is performed by comparing the corresponding characteristics curves of the devices. The parameters defined for the study include the output resistance of the devices, the channel length modulation factor, the leakage current, and the propagation delay time. Additionally, we study these effects by comparing the responses of a current mirror circuit implemented with stacked transistors against the current mirror version implemented using single devices. Moreover, a study of the physical implementation of arrays of stacked devices is presented.

In [2] we describe the development of a CAD tool to accelerate the layout implementation of two analog structures: the differential pair and an array of stacked devices. This CAD tool facilitates to circuit designers the analysis, test, characterization, and optimization of their designs. In the present paper, we exploit that CAD tool to parametrically generate multiple layout versions of stacked devices, analyzing how different implementations affect the performance of a specific circuit. Parameters considered for the implementation of the layouts are: the number of transistors in the array, the dimension of the devices, the number of fingers of each component and the selection of one of the two possible layout topologies referred before. We also compare the different implementations by performing an automated parasitic extraction process over each layout. This analysis allows a comparison of test performance and the identification of tradeoffs between the different stacked structures once they are implemented in layout. The technology used in this paper is a nanometric Intel process smaller than 100nm.

This work was supported in part by CONACYT (*Consejo Nacional de Ciencia y Tecnología*, Mexican Government) through a scholarship granted to I. Lomelí-Illescas and by Intel Corporation.

II. Effects on I/V Characteristics

As mentioned before, one of the main challenges in current nano-scale technology processes is the implementation of high performance analog circuits using technologies optimized for digital devices. A key limitation lies in the lack of flexibility to define the transistors' width and length. The width is limited to a set of discrete values, while transistor's length is normally limited to one single fixed value. Extensive research has been done to accurately describe the relationship between transistor dimensions and drain current in saturation region; e.g, in [3] they propose:

$$I_d = I_{dsat} = \frac{\mu_0}{[1 + U_0(V_{GS} - V_T)]} \cdot \frac{C_{ox}\frac{W}{L}(V_{GS} - V_T)^2}{2aK} \quad (1)$$

which includes short channel effects on CMOS devices. From (1) it is seen that I_d is still proportional to the ratio between the transistor's width and length. Since these dimensions are limited to discrete values in current nano-scale technologies, the options to achieve a specific analog performance are more limited. Let W_{min} denote the minimum feasible transistor width and WL_{min} the ratio between W_{min} and the fixed L associated to the technology. For a specific bias condition, the corresponding value of I_d is denoted as I_{dmin}. In Fig. 1, the output characteristic curves for devices with different feasible widths are shown, confirming limited discrete biasing currents. These curves where obtained from simulating an NMOS transistor in the above mentioned nanometric technology process. One of the solutions to increase the variety of biasing currents is by using transistors placed in stack. A stack array of N transistors is equivalent to a single transistor with N times its length [4]. With this it is possible to change not only the value of the transistor's width but also its length, or at least emulate this variation. The options for the selection of the width and length sizes are still discrete;

however, this will anyway increase the number of design options to achieve a desired performance for analog circuits.

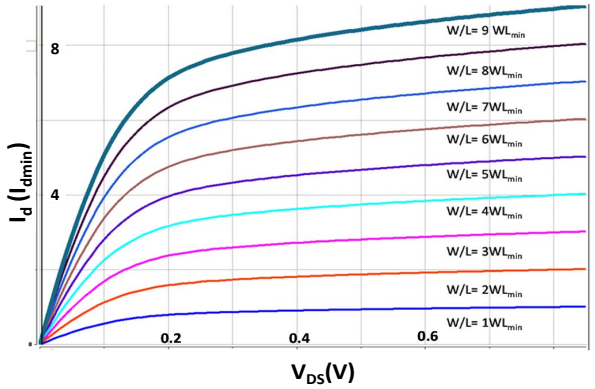

Fig. 1. Curves of an NMOS for multiple widths (W) and fixed length (L).

A. Channel Length Modulation

Ideally, in saturation region, CMOS transistors should behave as ideal voltage-controlled current sources. For a given V_{GS}, I_d should be constant and independent of V_{DS}. However, the effective channel length is actually modulated by V_{DS}, or as in current FinFETs technologies also by V_{GS} [5],

$$I_d = I_{sat}\,(1 + \lambda(V_{GS}) \cdot V_{DS}) \qquad (2)$$

where λ is the channel length modulation factor and is proportional to the inverse of the channel length. This factor typically increases for small devices. In Fig. 2, a comparison of the I_d curves of CMOS transistors for different W/L ratios and for different number of stack devices is shown. In this figure we can observe three different groups of curves; for each group the W/L ratio is the same, but the values of W and the number of stacked devices change. The slope values of these curves and the I_d value in the saturation region are shown in Table I; these values are normalized to the values obtained from a transistor of WL_{min} size; since the inverse of this slope represents the output resistance of the transistor, it is seen from Table I that the effect of the channel-length modulation factor is less important for long-channel transistors than for short-channel transistors. In addition, we can notice that I_d for specific bias conditions is higher when the number of stack devices is larger. The use of stack devices helps to obtain a better output resistance and smaller losses in the I_d current. However, this improvement becomes less significant as we continue increasing the number of devices in stack.

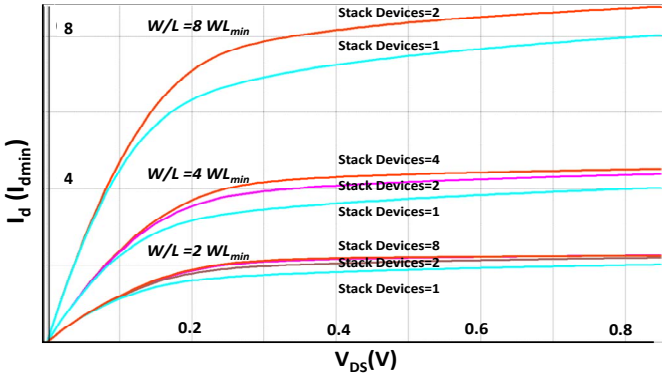

Fig. 2. Characteristics curves of a NMOS for multiple W and L values.

TABLE I. SLOPES FOR DIFFERENT W/L RATIOS

Width (W_{min})	Number of stack devices	W/L (WL_{min})	I_d (I_{dmin})	Slope ($Slope_{min}$)
16	2	8	2.09	12.00
8	1	8	1.56	1.61
16	4	4	0.93	0.80
8	2	4	0.88	0.61
4	1	4	0.78	0.43
16	8	2	0.47	0.40
8	4	2	0.46	0.31
4	2	2	0.44	0.21
2	1	2	0.39	0.15

B. Leakage

With the continuous scaling of CMOS devices, leakage current is becoming a major contributor to the total power consumption in a system. Many proposals have been developed to reduce its impact. In general, stacked devices have smaller leakage than the sum of the leakages consumed by all the devices, individually. This is often referred as the stack effect: the total leakage current of cascade transistors chain decreases as the number of stacked transistor increases [6]. In modern deep sub-micron devices, the threshold voltage may decrease for longer channels due to the reverse short channel effect. Therefore, leakage reduction is less effective, but is still a commonly used technique. In Table II, the values of leakage current for a different number of stacked devices are summarized. These values are also normalized taking as reference the leakage current of a WL_{min} transistor: I_{leak_min}. It is shown that in off state the subthreshold current is significantly smaller than for a single device.

C. Current Mirror

In this subsection, the effects of using arrays of stacked transistors in a commonly used analog circuit are analyzed. As an example, the circuit studied here is a current mirror. Some important features of the current mirror are the following:

a) Relatively high output resistance to keep the output current constant regardless of the load conditions.

b) Relatively low input resistance to keep the input current constant regardless of drive conditions.

c) Output current linearly mirrored: $i_o = A_i i_i$.

In Fig. 3a the schematic diagram of a current mirror is presented. As it is well known, in a simple current mirror, if we assume that $V_{DS2} > V_{GS}–V_{T2}$, then i_o can be obtained from

$$\frac{i_o}{i_i} = \left(\frac{L_1 W_2}{L_2 W_1}\right)\left(\frac{V_{Gs} - V_{T2}}{V_{Gs} - V_{T1}}\right)^2 \left[\frac{1 - \lambda_{VDS2}}{1 - \lambda_{VDS1}}\right]\left(\frac{K_2}{K_1}\right) \qquad (3)$$

If the transistors are ideally matched, then $K_1 = K_2$ and $V_{T1} = V_{T2}$, and if $V_{DS1} = V_{DS2}$, then

$$\frac{i_o}{i_i} = \left(\frac{L_1 W_2}{L_2 W_1}\right) \qquad (4)$$

Therefore, the sources of error are the difference between V_{DS1} and V_{DS2} and the mismatch between M1 and M2 [7].

In Fig. 3b, the output resistance of a simple current mirror is shown, varying the number of transistors placed in stack. It is clearly seen from Fig. 3b that channel modulation effects are reduced when more transistors in stack are used. In Table II, the normalized results, including the matching percentage

between input and output currents of the circuit, are summarized. From here it is confirmed that a better current matching is achieved as we increase the number of stacked devices.

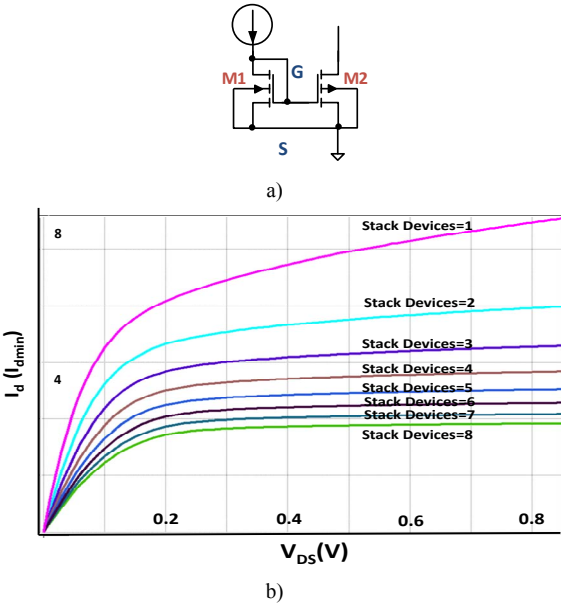

Fig. 3. Simple current mirror: a) schematic diagram; b) curves for the output resistance varying the number of stack devices.

TABLE II. RESULTS VARYING THE NUMBER OF STACKED DEVICES

Stack devices	Single device				Current mirror	
	I_d (I_{dmin})	Slope ($Slope_{min}$)	Leakage current ($I_{leak\ min}$)	Delay time ($Delay_{min}$)	% matching	Slope ($Slope_{min}$)
1	1.00	1.00	1.00	1.00	92.094	1.00
2	0.56	0.38	0.35	2.17	96.315	0.40
3	0.39	0.21	0.11	2.84	97.484	0.24
4	0.29	0.13	0.09	3.40	97.840	0.17
5	0.24	0.09	0.08	3.90	98.380	0.12
6	0.17	0.07	0.07	4.41	99.107	0.14
7	0.15	0.05	0.06	4.90	99.310	0.10
8	0.12	0.04	0.04	5.56	99.780	0.07

D. Delay Time

Despite the advantages of using stacked devices, some negative effects need to be considered. There is a tradeoff between power and delay in the propagation of signals. Due to the input load requirement and due to the stacking of devices, the drive current of a forced-stack gate will be lower, resulting in an increased delay [8]. In Table II, the delay results for the propagation of a pulse at the input of an array of stacked devices are presented, taking as reference the propagation delay through a single WL_{min} device. It is confirmed from Table II that the delay time increases as more elements are included in the array.

III. STACKED DEVICE LAYOUT TOPOLOGIES

Three different topologies for implementing stacked device layout are presented in [2]. Here we consider the first two of them, as follows:

a) Topology A or one shared diffusion. The transistors are divided in fingers and are placed one next to each other. The source of one transistor is shared with the diffusion of the next one. All fingers of one device are placed next to all the fingers of the next one.

b) Topology B or interdigitated layout implementation. This topology is used when the transistors are divided in at least two fingers; one finger of each transistor is placed next to each other, starting from that one on the "top" of the array and continuing until one finger of the transistor that is on the "bottom". Then, the order in which the transistor fingers are placed is inverted; this process is repeated as many times as the number of transistor fingers of the transistor is completed.

IV. IMPLICATIONS OF LAYOUT TOPOLOGIES

Here we analyze both layout topologies described above. Our analysis is based on two criteria: a) interconnection complexity and b) parasitic elements values and area.

A. Interconnection Complexity

For argument sake, we consider an array of four-stacked devices. If the number of fingers per transistor is one, both layout implementations would be equal, including their routing. When the number of fingers is larger than two, the routing of the topologies will be different: the lengths of the interconnections metals are longer in the case of the interdigitated layout topology. Furthermore, if the number of fingers per transistor increases, the difference in length also increases. Similarly, if more transistors are added to the array, the length of the interconnections on the interdigitated layout topology will be much longer. In addition, more routing tracks will be needed. This is illustrated in Fig. 4, where we can also see that in the case of one shared diffusion topology, the metals used for routing can share tracks, while in the case of the interdigitated layout each metal need its own track. As more transistors are included in the array, more routing tracks will be needed; this will increase the area that is required for the layout implementations. As a solution some of the signals could be routed using higher metal layers. This will reduce the area for the implementation, but the complexity in the layout design and the mismatch between the devices increases

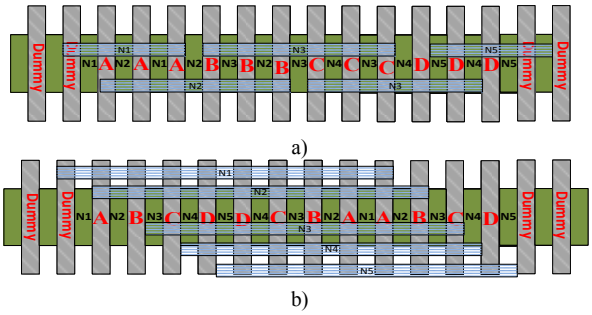

Fig. 4. Topologies for the layout implementation of an array of stacked devices: a) one shared diffusion; b)interdigitated layout.

B. Parasitic Elements and Area

In this subsection, layout implementations for arrays of stacked devices are compared in terms of their parasitic effects. For this we consider circuit arrays from two to four devices. The width of each device is six times W_{min}. Each of these devices could be divided in one, two or three fingers. To compare all the possible implementations, the input capacitance (in_C) and the variation of the cross capacitance (C_C, the capacitance from the net to the rest of layout elements) between the interconnections nodes are obtained. These nodes were chosen to analyze how the topologies and metal interconnections affect the current flows through this path due to differences between the capacitance associated to

them. In addition, the sum of the capacitances of all these nodes and the gate node to VSS (*Cvss*) are obtained; this capacitance is normally associated to leakage effect. This value could indicate if some of the implementations have a better performance in this regard. Finally, the area of these implementations is obtained. Results for all these implementations are summarized in Tables III-V. The results are normalized with respect to the values obtained from the layout implementation of a single device of WL_{min} dimensions.

When the number of fingers is one, both implementations are equal and these results are the same (see Table III). From Table IV we can notice that for the shared diffusion implementation, the variation on the parasitic values are smaller than for the interdigitated layout; this is due to the variations in the lengths of the metals used to interconnect the fingers of each transistor. Interdigitated implementations have a better device matching than the shared diffusion implementations. However, the metal interconnections of this topology does not show good matching between them.

Tables III-V confirm that as more elements are included in the layout, more parasitics are generated, deteriorating circuits' performance. Naturally, area increases as more elements are included, but the area tends to be larger when transistors are divided in fingers (due to the area required for routing signals; actually when two fingers are used the area increases due to the use of middle dummy devices). However, the values of the parasitic elements are smaller when a single device is used than when it is divided in fingers. Furthermore, the layout connectivity is less complex. Designers should consider to use the minimum number of fingers to reduce the generation of parasitic elements and the complexity on the layout implementation. In Fig. 5 we present a comparison of the required layout area when different number of stacked devices are used, including also the curve of the slope values. It is seen that as the number of stacked devices increases, the improvement on the output resistance is less significant, but the required area increases almost linearly.

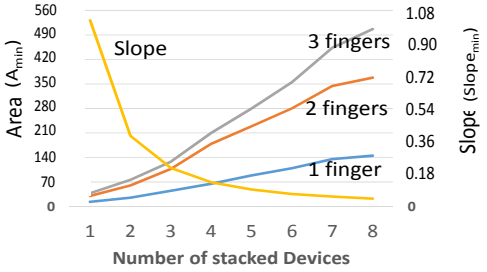

Fig. 5. Comparison of the layout area and output slope for different number of stacked transistors varying the number of fingers.

V. Conclusions

In this paper, relevant implications on the use of stacked transistors topology were parametrically analyzed. The use of stacked transistors increases the number of options for the design of modern analog circuits. It also offers additional advantages: significant reduction of leakage current (stack effect); increase of the output resistance with respect to a single device, which is especially useful for more complex structures, as it was illustrated with the use of current mirrors. However, they also have some drawbacks, such as increased propagation delays. In terms of layout topologies, as more elements are included: the values of the parasitic elements increase, reducing the maximum operating frequency of the circuit; the area of the layout increases and the complexity of its implementation also increases. Our methodology will help designers to consider all these tradeoffs for the optimal implementation of their designs before expensive physical testing on silicon.

TABLE III. Layout Results using One Finger per Transistor

Stack devices	Share diffusion / Interdigitated layout			
	Input Cap (I_{Cmin})	Cap Var. (%)	*Cvss* (*Cvss_min*)	Area (A_{min})
2	6.00	1.15	46	16.36
3	6.32	5.28	57	21.14
4	6.36	1.52	97	31.64

TABLE IV. Layout Results using Two Fingers per Transistor

Stack devices	Share diffusion				Interdigitated layout			
	Input Cap (in_{Cmin})	Cap Var. (%)	*Cvss* (*Cvss_min*)	Area (A_{min})	Input Cap (in_{Cmin})	Cap Var. (%)	*Cvss* (*Cvss_min*)	Area (A_{min})
2	6.87	27.09	18.26	30.36	6.91	53.88	21.68	35.36
3	6.55	25.40	29.42	50.71	7.26	51.59	33.05	63.57
4	6.87	10.41	42.47	83.21	6.51	19.76	51.95	116.07

TABLE V. Layout Results using Three Fingers per Transistor

Stack devices	Share diffusion				Interdigitated layout			
	Input Cap (in_{Cmin})	Cap Var. (%)	*Cvss* (*Cvss_min*)	Area (A_{min})	Input Cap (in_{Cmin})	Cap Var. (%)	*Cvss* (*Cvss_min*)	Area (A_{min})
2	9.47	12.75	15.68	25.00	10.38	20.55	13.32	25.76
3	9.63	14.29	18.79	52.50	10.42	16.80	19.11	45.71
4	9.55	11.03	30.05	53.21	11.29	13.07	31.00	66.43

References

[1] L. L. Lewyn, T. Ytterdal, C. Wulff, and K. Martin, "Analog circuit design in nanoscale CMOS technologies," *Proceedings of the IEEE*, vol. 97, no. 10, pp. 1687-1714, Oct. 2009.

[2] I. Lomelí-Illescas, S. A. Solis-Bustos, V. Martínez, and J. E. Rayas-Sánchez, "Synthesis tool for automatic layout generation of common analog structures," in *IEEE Andean Council Int. Conf. (ANDESCON 2016)*, Arequipa, Peru, Oct. 2016.

[3] B. J. Sheu, D. L. Scharfetter, P. K. Ko and M. C. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," in *IEEE Journal of Solid-State Circuits*, vol. 22, no. 4, pp. 558-566, Aug. 1987.

[4] D. Kong, D. Seo and S. M. Lee, "Analysis and reduction of nonidealities in stacked-transistor current sources," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. PP, no. 99, pp. 1-5, Aug. 2016.

[5] Nariai and A. Hiroki, "Gate voltage dependence of channel length modulation for 14nm FinFETs," *2016 IEEE International Meeting for Future of Electron Devices*, Kansai (IMFEDK), Kyoto, 2016, pp. 1-2.

[6] N. Saxena and S. Soni, "Leakage current reduction in CMOS circuits using stacking effect," *Int. Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 2, no. 11, pp. 213-216, Nov. 2013.

[7] R. L. Geiger, P. E. Allen, and N. R. Strader, *VLSI Design Techniques for Analog and Digital Circuit*s. New York: Mc-Graw Hill, 1990.

[8] S. Narendra, S. Borkar, V. De, D. Antoniadis and A. Chandrakasan, "Scaling of stack effect and its application for leakage reduction," in *Int. Symp. on Low Power Electronics and Design*, Huntington Beach, CA, pp. 195-200, Aug. 2001.