Leakage Reduction in Stacked Sub-10nm Double-Gate MOSFETs

Woo-Suhl Cho and Kaushik Roy
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47906, USA
Email: {cho68, kaushik}@purdue.edu

***Abstract*—In this paper, the effectiveness of transistor stacking (or supply-gating) to reduce the leakage in the standby-mode of operation of sub-10nm double-gate MOSFETs is investigated. For that purpose, device parameters such as symmetric/asymmetric gate-to-source/drain underlap and body thickness are optimized to improve the ON-state current to the OFF-state current ratio. The optimized devices are then used in circuit simulation to analyze the dependence of each major leakage source (direct source-to-drain tunneling, thermionic, and gate oxide leakage currents) on the device geometry (t_{si} and symmetry in L_{UN}) and input vectors for two- and three-stacked transistors. The analysis shows that supply-gating is effective in reducing direct source-to-drain current as well as thermionic leakage in the stand-by mode of operation for sub-10nm technology.**

Keywords—Direct source-to-drain tunneling, Sub-10nm double-gate MOSFETs, Stacking, Supply-gating

I. INTRODUCTION

Multi-gate transistor such as a double-gate MOSFET (DGFET) is inevitable for sub-10nm technology nodes due to its high immunity to the short channel effects (SCE). However, there is a need for the modification in device design at this deeply-scaled regime since device characteristics are also affected by other leakage mechanism such as direct source-to-drain tunneling (DSDT) [1-3]. Narrow channel potential barrier of a sub-10nm gate length transistor gives rise to quantum mechanical tunneling of electrons from source to drain (I_{DSDT}). The total OFF-state leakage current then becomes the sum of I_{DSDT}, gate oxide leakage current (I_G), and thermionic current (I_{THERM}) over the channel potential barrier as shown in Fig. 1 (a). One of the effective ways to reduce DSDT is to increase the effective channel length by introducing gate/source or gate/drain underlap without increasing gate length (L_G) [4]. Larger channel length also mitigates SCE, and leads to the increase in ON-state current (I_{ON}) to the OFF-state current (I_{OFF}) ratio compared to a corresponding non-underlapped sub-10nm device [2]. The use of gate-to-source/drain underlap also reduces I_G since edge direct tunneling (EDT) current between gate and source/drain overlaps become negligible.

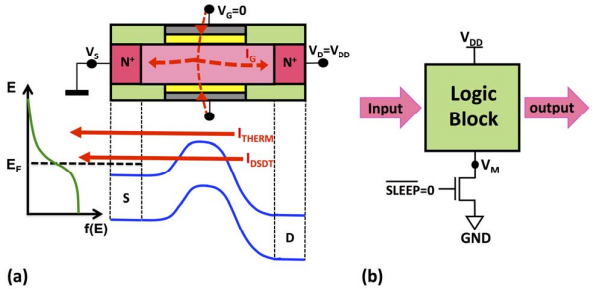

Fig. 1. Major leakage current sources in a sub-10nm n-type DGFET in the OFF-state. (b) Schematic of supply-gating for logic.

On the other hand, in order to reduce the leakage power of the circuit in the standby-mode of operation, supply-gating technique shown in Fig. 1 (b) has been widely used [5-8]. In Fig. 1 (b), an extra NMOS transistor inserted between the ground (GND) and the logic block creates series-connected NMOS stacks with the pull-down transistors in the logic block. During the sleep mode of operation, this gating transistor turns off, and makes the virtual GND node (V_M) to be higher than GND. Pull-down transistors with an input of logic 0 in the logic block then operates at negative gate-to-source bias (V_{GS}) due to positive value of V_M. In sub-100nm technologies, substantial leakage reduction in the standby-mode was obtained [5-8] since I_{OFF} is exponentially dependent on V_{GS}. The technique of using stacked transistors to reduce the leakage in standby-mode ("stacking effect") needs to be reinvestigated to evaluate its effectiveness for deeply-scaled technology.

To that effect, first, we optimized DGFETs using symmetric (L_{UN}) and asymmetric underlaps on source ($L_{UN,S}$) and drain ($L_{UN,D}$) sides to improve I_{ON}/I_{OFF}. We then performed detailed device-circuit mixed-mode analysis using the optimized devices in order to understand the impact of supply-gating in leakage reduction. The rest of the paper is organized as follows. Section II describes simulated device structure and our simulation framework. In Section III, we explore the design parameters to optimize the devices. Section IV analyzes the effectiveness of transistor stacking and hence,

This work was supported by the DARPA's Power Efficiency Revolution For Embedded Computing Technologies (PERFECT) Program.

978-1-4673-7860-4/15/$31.00 ©2015 IEEE

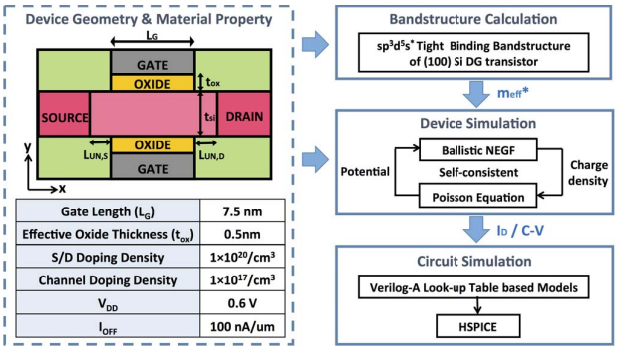

Fig. 2. The baseline device structure with its parameters and the simulation framework.

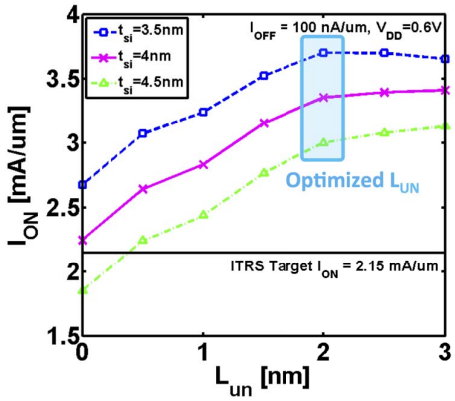

Fig. 3. I_{ON} under iso-I_{OFF} as a function of symmetric L_{UN} for different t_{si}.

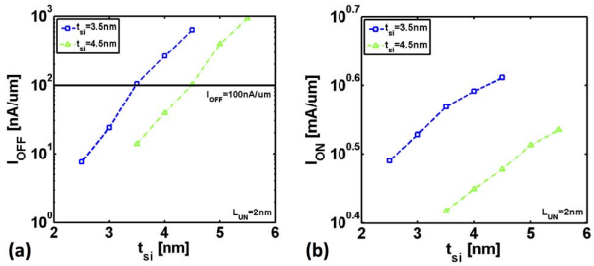

Fig. 4. (a) I_{OFF}, and **(b)** I_{ON} changes with the variation in t_{si}.

supply-gating, in leakage reduction. Finally, Section V derives conclusions from the results of our analysis.

II. Device Structure and Simulation Framework

The Si DGFET structure and the modeling framework used for the simulations and analysis in this work is shown in Fig. 2. The baseline device is designed following International Technology Roadmap for Semiconductor (ITRS), and has L_G=7.5nm and t_{si}=4.5nm. It also has the effective oxide thickness (t_{ox}) of 0.5nm (HfO_2+SiO_2 layers), and SiO_2 is used for spacers. [9]. The channel and source/drain doping concentrations are 10^{17}/cm^3 and 10^{20}/cm^3, respectively, and the source/drain dopant concentration is assumed to drop 1 decade/nm toward the channel region with a Gaussian profile. The devices are then optimized with symmetric/asymmetric gate-to-source/drain underlaps, and body thickness (described in section III). Note, all the devices are designed with iso-I_{OFF} of 100nA/um at supply voltage (V_{DD}) of 0.6V [9] by tuning their gate work-function.

In order to understand the impact of supply-gating on leakage reduction, we used the simulation framework based on physics-based device models coupled with the circuit equations [2], [10]. To capture the quantum mechanical effects such as quantum confinement and DSDT, which are critical in scaled technologies, we first extracted the effective mass from sp^{3}d^{5}s*tight-binding (TB) band-structure [11] of (100) Si thin film for different body thickness. Poisson equation is then solved self-consistently with the ballistic Non-equilibrium Greens function (NEGF) [12] using the extracted effective masses and following the mode space approach [13]. The resulting characteristics – I_{THERM}, I_{DSDT}, I_G, and capacitance – are used in look-up table based Verilog-A models to perform circuit simulations in HSPICE. Equipartition of I_G between source and drain [14] is assumed.

III. Device Optimization

In this section, we optimize device parameters such as symmetric/asymmetric gate-to-source/drain underlaps and body thickness to obtain high I_{ON}/I_{OFF} using the device structure and the simulation framework discussed in the previous section. The sensitivity of I_{ON} and I_{OFF} to the variation in the body thickness is also considered to optimize the body thickness.

A. Symmetrically Underlapped Devices

Gate-to-source/drain underlap (L_{UN}) can be used to improve the sub-threshold characteristics of the device for sub-10nm technology. This is because it gives wider channel potential barrier, which reduces DSDT, and mitigates SCE [2-3]. The overlapped capacitance, and I_G also improve as a by-product of underlapping. As a result, an increase in I_{ON}/I_{OFF} (I_{ON} under iso-I_{OFF}) can be achieved with an increase in L_{UN} as shown in Fig. 3. The symmetric L_{UN} of 2nm, which gives the highest I_{ON}/I_{OFF}, is chosen from Fig. 3 as the optimal device. Note, increasing L_{UN} comes at a cost of increased footprint, and L_{UN} beyond 2nm lead to reduction in I_{ON}/I_{OFF} because of increased channel resistance. On the other hand, lower t_{si} further increases I_{ON} under iso-I_{OFF} since it improves the sub-threshold characteristics [2]. However, lower t_{si} is associated with variability issues arising from quantum confinement [15]. Therefore, we considered the sensitivity of I_{ON} and I_{OFF} to the variation in the body thickness to optimize the body thickness. Fig. 4 compares I_{OFF} and I_{ON} changes with the variation in t_{si} for two devices (t_{si}=3.5nm and t_{si}=4.5nm). It is shown that t_{si} can be reduced to 3.5nm without having much of an impact on variation compared to t_{si} of 4.5nm.

B. Asymmetrically Underlapped Devices

Asymmetrically underlapped devices can be used in 6T SRAM to mitigate the read-write design conflict [16].

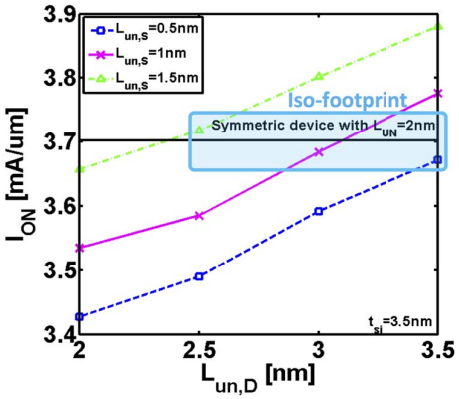

Fig. 5. I_{ON} under iso-I_{OFF} of the devices with different asymmetric underlaps and t_{si}=3.5nm.

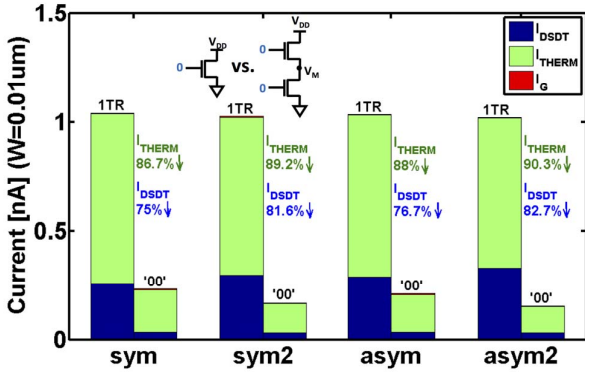

Fig. 6. Reduction in each leakage components of two-stacked OFF transistors.

TABLE I
THE PARAMETERS OF OPTIMIZED SYMMETRIC AND ASYMMETRIC DEVICES

Device	t_{si}	$L_{UN,S}$	$L_{UN,D}$
Sym1	3.5nm	2nm	2nm
Sym2	4.5nm	2nm	2nm
Asym1	3.5nm	1.5nm	2.5nm
Asym2	4.5nm	1.5nm	2.5nm

TABLE II
THE IMPACT OF STACKING ON SHORT CHANNEL EFFECTS

Device	DIBL	ΔV_{TH}
Sym1	98 [mV/V]	2.3 [mV]
Sym2	136 [mV/V]	4.6 [mV]
Asym1	108 [mV/V[	2.8 [mV]
Asym2	149 [mV/V]	5.7 [mV]

Therefore, in this work, for possible future memory application, we optimize the device using asymmetric underlaps on the source and the drain sides. Smaller underlap is used on the source side compared to the drain side to obtain higher I_{ON}/I_{OFF} [16]. The circuit behavior of asymmetric devices are compared to the corresponding symmetric devices in the next section.

Fig. 5 shows I_{ON} under iso-I_{OFF} of the devices with different asymmetric underlaps. I_{ON}/I_{OFF} increases as the underlap increases since better sub-threshold characteristics are obtained for devices with larger effective channel length as discussed in the last section. Therefore, asymmetric devices with the same footprint as the optimized symmetric device (with L_{UN}=2nm) have similar level of I_{ON}. Since slight increase in I_{ON}/I_{OFF} also comes at a cost of larger footprint, we chose these iso-footprint asymmetric devices as the optimal devices. Table 1 shows our four selected devices used for the circuit simulation in the next section.

IV. IMPACT OF SUPPLY-GATING IN SUB-10NM TECHNOLOGY

In this section, we analyze the effectiveness of supply-gating to improve leakage in the standby-mode of operation for our optimized devices. We also compare the dependence of each leakage source on device geometry (t_{si} and symmetry in L_{UN}) and input vectors in stacked transistors.

A. Leakage Reduction in Stacked Transistors

In order to analyze the impact of supply-gating, we consider the "stacking effect" in two-stacked (or series connected) NMOS transistors. Fig. 6 compares major leakage sources of a single NMOS transistor in OFF-state to the currents flowing through a stack of two OFF transistors. It is observed that I_{DSDT} is much smaller than I_{THERM} in our devices since they are optimized with gate-to-source/drain underlaps to reduce DSDT. Also, underlapping gives very small I_G compared to other leakage components since EDT current becomes negligible. It is also shown in Fig. 6 that stacking is effective in reducing both I_{DSDT} and I_{THERM}. When both transistors in the stack (inset in Fig. 6) are turned off, leakage current flowing at the bottom transistor leads to positive value of intermediate node voltage ($V_M > 0$). The top transistor then operates at negative V_{GS}, and the corresponding sub-threshold current flowing through the stack reduces exponentially. In addition, decreased drain to source bias (V_{DS}) of the top transistor reduced drain-induced barrier lowering (DIBL) leading to further leakage reduction. As a result, the reduction in the total leakage current is obtained in transistor stacks.

On the other hand, it is also observed in Fig. 6 that stacking is more effective in reducing the leakage for devices with thicker body. This is because they have higher DIBL as shown in Table II due to weaker gate control. Larger channel resistance increases V_M and lowers V_{GS}, leading to lower sub-threshold current for the top transistor in a stack. Note, ΔV_{TH}, which is the threshold voltage shift of the top transistor in the stack compared to the single transistor, is higher for the devices with thicker body. Hence, more leakage reduction can be achieved for devices with thicker body.

B. Leakage for Different Input Vectors

Fig. 7 shows each leakage source and the total leakage current in two series connected transistors with different input vectors ('00', '10', and '01'). It is observed in Fig. 7 (a) that

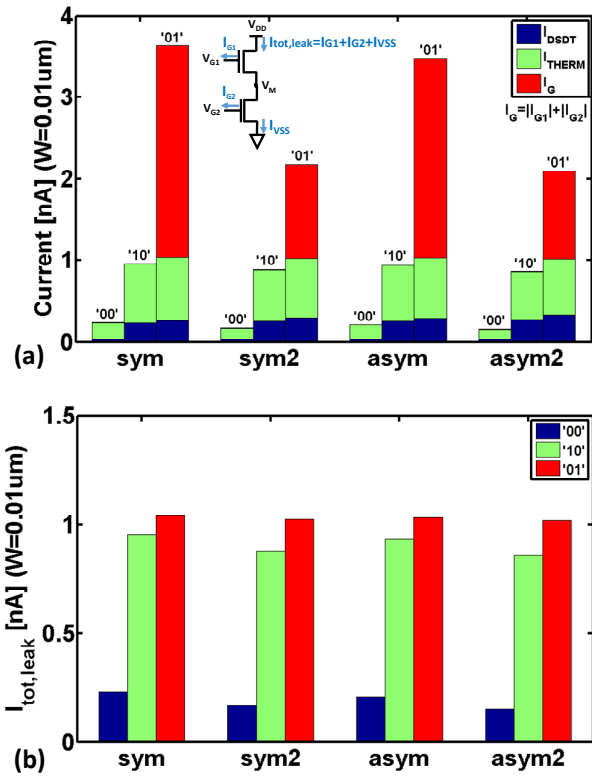

Fig. 7. (a) Major leakage components, and (b) total leakage current of two-stacked optimized transistors with different input vectors.

for the input vector '00', I_{THERM} and I_{DSDT} are much lower compared to other input vectors due to the negative V_{GS} (V_M>0) operation of the top transistor. It also leads to the reduction in the total leakage current, which is the sum of all the currents going to the ground (or, all the current coming out of the supply), as shown in Fig. 7 (b). However, I_G shows different dependence on input vectors compared to other leakage components. In our optimized devices with underlapping, EDT current is negligible and only tunneling current between gate and the inverted channel dominates the gate current of ON transistor. Therefore, input state '01' has the highest I_G since the bottom transistor operates in strong inversion (V_{GS}=V_{DD}). On the other hand, with '10' as an input vector, very small gate-to-channel tunneling current flows as in '00'. This is because the top transistor in the stack operates at weak inversion (V_{GS}=V_{TH}) since V_M rises to V_{DD}-V_{TH}, where V_{TH} is the transistor threshold voltage.

V. Conclusion

In this paper, we investigated the effectiveness of supply-gating to improve leakage in the standby-mode of operation for sub-10nm technology where severe SCE along with new leakage mechanisms such as DSDT exists. For that purpose, first, we optimized sub-10nm DGFETs using symmetric/asymmetric underlap and considering the sensitivity of I_{ON} and I_{OFF} to the variation in the body thickness. Major leakage current sources of the optimized devices are estimated using quantum device simulation. The resulting device characteristics are then used in circuit simulation to analyze the dependence of each leakage source on the device geometry (t_{si} and symmetry in L_{UN}) and the applied input vectors for two- (and three)-stacked transistors. The analysis shows that supply-gating is effective to reduce I_{THERM} as well as I_{DSDT} in sub-10nm devices.

References

[1] J. Wang and M. Lundstrom, "Does source-to-drain tunneling limit the ultimate scaling of MOSFETs," in *IEDM Tech. Digest,* 2002, pp. 707-710.

[2] W. S. Cho, S. K. Gupta, and K. Roy, "Device-Circuit Analysis of Double-Gate MOSFETs and Schottky-Barrier FETs: A Comparison Study for Sub-10nm Technologies," *IEEE Trans. Electron Devices*, vol. 61, no.12, pp. 4025-4031, Dec. 2014.

[3] W. S. Cho and K. Roy, "The effects of direct source-to-drain tunneling and variation in the bodythickness on (100) and (110) sub-10nm Si double-gate transistors," *IEEE Electron Device Lett.*, vol. 36, no. 5, pp. 427-429, May 2015.

[4] S. Balasubramanian, L. Chang, B. Nikolic, and T. J. King, "Circuit-performance implications for double-gate MOSFET scaling below 25nm," in *Proc. Silicon Nanoelectronics Workshop*, pp.16-17, June 2003.

[5] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor stacks," in *Proc. Int. Symp. Low Power Electronics and Design*, 1998, pp. 239-244.

[6] M. C. Johnson, D. Somasekhar, and K. Roy, "Models and Algorithms for Bounds on Leakage in CMOS Circuits," in *IEEE Trans. Comp.-Aid. Des. Int. Cir. and Syst.*, vol. 18, no.6, pp. 714-725, JUNE 1999.

[7] Y. Ye, S. Borkar, and V. De, "New technique for standby leakage reduction in high-performance circuits," in *Symp. VLSI Circuits Dig. Tech. Papers*, 1998, pp. 40-41.

[8] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of the IEEE*, vol. 91, no.2, pp. 305-327, Feb. 2003.

[9] International Technology roadmap for Semiconductors (ITRS) 2012. [Online]. Available: http://public.itrs.net

[10] S. Gupta, W. Cho, A. A. Goud, K. Yogendra, and K. Roy, "Design Space Exploration of FinFETs in Sub-10nm Technologies for Energy-Efficient Near-Threshold Circuits," in *DRC*, pp.117-118, 2013.

[11] G. Klimeck, F. Oyafuso, T. B. Boykin, R. C. Bowen, and P. Allmen, "Development of a Nanoelectronic 3-D (NEMO 3-D) Simulator for Multimillion Atom Simulations and Its Application to Alloyed Quantum Dots" (INVITED), *Computer Modeling in Engineering and Science*, vol. 3, no. 5, pp 601-642, 2002.

[12] S. Datta, "Nanoscale device modeling: The Green's function method," *Superlatt. Microstruct.*, vol. 28, pp. 253-278, 2000.

[13] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, "Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches," *J. App. Phys.* 92(7), 2002.

[14] C. Choi, K. Nam, Z. Yu, and R. W. Dutton, "Impact of gate direct tunneling current on circuit performance: a simulation study," *IEEE Trans. Electron Device*, vol. 48, pp.2823-2829, Dec. 2001.

[15] S. Xiong and J. Bokor, "Sensitivity of double-gate and finfet devices to process variations," *IEEE Trans. Electron Devices*, vol. 50, no.11, pp. 2255-2261, Nov. 2003.

[16] A. Goel, S. K. Gupta, and K. Roy, "Asymmetric drain spacer extension (ADSE) FinFETs for low-power and robust SRAMs," *IEEE Trans. Electron Devices*, vol. 58, no.2, pp. 296-308, Feb. 2011.