Journal of VLSI Design Tools and Technology
Volume 2, Issue (1-3), Compilation 2012

STM JOURNALS

Detailed Study of Working and Applications of FinFET Technology and Its Adaptability in Current Era

Manan Sheth[1]*, Shivang Bakliwal[1], Shival Trivedi[1], Dhaval Shah[2]
[1]Electronics and Communication Department, Nirma University, India
[2]L. D. College of Engineering, India

ABSTRACT

The CMOS technology is following Moore's law since almost 25 years. The scaling of CMOS transistor has increased packaging density, speed and decreased power dissipation in the integrated circuits. However, CMOS dimension when scaled to nanometer dimension, many physical barriers arise. In sub-100 nm scale, MOSFET has new variants as SOI implementation and double gate.

Keywords: *Silicon on insulator (SOI), short channel effect (SCE) [2], buried oxide (BOX)*

* ***Author for Correspondence*** E-mail: manntsheth@gmail.com

1. INTRODUCTION

During the shrinking of conventional single-gate MOSFET, the charge carrier mobility is changed and the channel length becomes of the same dimension as depletion layer depth. The gate voltage is not capable of handling switching of MOSFET when drain-induced barrier lowering occurs. So, the control of gate voltage over channel operation is diminished when the scaling down of MOSFET happens. The depletion layer at drain and source covers charges and behaves as a capacitor and it disables the high-frequency operation. There are scaling limits imposed by the non-scalability of silicon energy band gap and thermal voltage. The non-scaling of silicon band gap energy leads to non-scaling of built-in potential, depletion width and short channel effect. The non-scaling of thermal voltage leads to sub-threshold non-scaling.

The physical dimensions are also limited by quantum mechanical tunneling currents that pass through various barriers in MOSFET. As the oxide thickness is scaled towards 1.5 nm, which corresponds to 2 to 3 layers of SiO_2 atoms, the oxide tunneling current induces gate leakage, and thereby increases standby power dissipation. Moving to multiple-gate MOSFETs might be the viable alternative to build MOSFETs with gate length $L_g < 50$ nm.

2. MULTIPLE-GATE MOSFET

Multiple-gate MOSFET can provide better control of the channel to the gates. Reduced short-channel effects, off-state leakage current, higher current drive capability, improved sub-threshold slope, reduced drain-to-body capacitance and higher carrier mobility can be obtained practically.

The short-channel effects can be suppressed without increasing the channel impurity concentration. Double-gate MOSFETs can be scaled to the shortest channel length possible for a given gate oxide thickness, because the bottom

gate can effectively screen the field penetration from the drain, and thus it can suppress the short channel effects. 60 mV/decade sub-threshold slope, scaling by silicon film thickness without high doping, setting of threshold voltage by gate work functions are the benefits achieved by double-gate MOSFET.

In double-gate MOSFET, the top and bottom gates can be driven together to obtain larger I_{on}/I_{off} ratio, or independently to allow for dynamic threshold voltage modulation. Double-gate MOSFET can be fabricated as a planar or non-planar design. Non-planar design of multiple gates is known as FinFET, where the conducting channel is wrapped in a thin silicon fin.

3. SILICON ON INSULATOR DESIGN

Silicon on Insulator **(**SOI) was originally invented for application in special environmental conditions, such as radiation-hardened or high-voltage integrated circuits [1].

Now, SOI has become a better way to make low-power and high-performance circuits. The buried oxide formed on the bulk silicon substrate has active MOS devices and circuits on the top of it. The parasitic capacitance of MOSFET is reduced due to use of buried oxide (BOX). Thus, the delays of digital CMOS circuits are decreased and the operating speed can be increased by controlling junction capacitance. Power-delay product of SOI CMOS circuits is also smaller because of less parasitic capacitance; also the leakage currents are reduced.

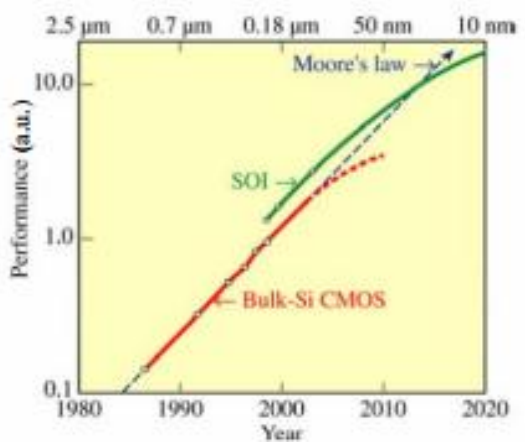

Fig. 1: *Growth of Bulk MOSFET and SOI MOSFET over Years and Performance Is in Arbitrary Unit* [1].

3.1. Classification of Silicon on Insulator MOSFETs

Silicon on Insulator-based MOSFETs is mainly divided into fully depleted and partially depleted devices [1]. A fully depleted device has the SOI layer much smaller than depletion region width and its potential is tightly controlled by the gate. For partially depleted device, the SOI layer is thicker than the maximum depletion width of the gate. The SOI device makes the manufacturing process more compatible than a traditional MOSFET.

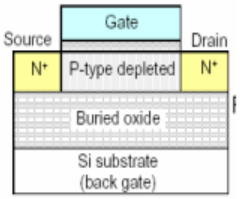

Fig. 2: *Structure of a Fully Depleted SOI MOSFET* [1].

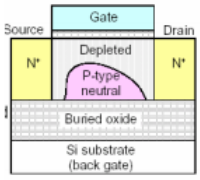

Fig. 3: *Structure of a Partially Depleted SOI MOSFET* [1].

4. FABRICATION METHOD OF FINFET [2]

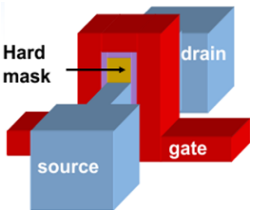

Fig. 4: *Structure of a Double-Gate MOSFET.*

The heart of a FinFET is fin which is considered as the body of MOSFET. The length of the fin from the source to the drain determines the effective channel length of the device. The heavily doped polysilicon film reduces the series resistance between the source and the drain. The gap between the source and the drain is reduced by dielectric spacers. On fabrication, the thick buried oxide of almost 400 nm thickness is layered on silicon substrate of 50 nm thickness. The use of SOI technology is to reduce parasitic capacitance. Sapphire or silicon dioxide is used as the insulator material as per the requirement of the application. A hard mask is provided over the fin for protection during etching process. After applying hard mask, the photo-resistive material is applied on the hard mask to define the shape and size of the device. After that, optical lithography or some other process like electron beam lithography is applied which is followed by the etching process which is to remove the hard mask and silicon fin from the oxide surface. After that, the gate material is deposited over the oxide which can be polysilicon or some other compound like titanium nitride or molybdenum. On the gate, the mask is applied and with the same process described, the etching of excess gate material is done. After the designing of the gate, the mask over it can be removed. The dielectric spacers of silicon nitride or silicon dioxide are then formed. After these processes, the mask is removed from the source and the drain parts for direct exposure and for better conduction, the source and drain are doped with different techniques like ion implantation or gas diffusion process.

Fig. 5: *After Depositing Si_3N_4 and SiO_2 Hard Mask, Si Fin Is Formed by Etching* [2]

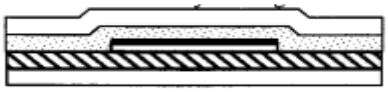

Fig. 6: *Phosphorus-Doped-Poly Si and SiO_2 Stacked Layer Deposited* [2]

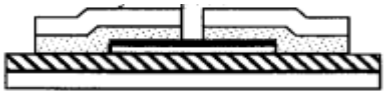

Fig. 7: *Source and Drain Were Etched while Si Fin Is Protected by the Hard Mask* [2]

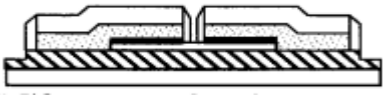

Fig. 8: *SiO_2 Spacers are Formed* [2].

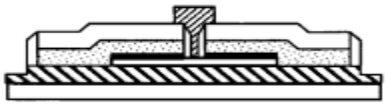

Fig. 9: *After Doping B-Doped SiGe, Gate Pattern Was Delineated* [2]

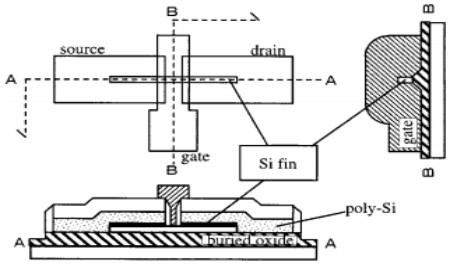

Fig. 10: *FinFET Typical Layout and Cross Sectional Structures* [2]

As another alternative of the multiple-gate MOSFET design with the single gate, multiple sources and drains are provided. With one source and drain, more than one channel can be joined.

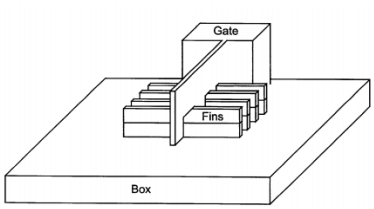

Fig. 11: *Multiple Sources and Drains with Single Gate* [3].

4.1. Design Parameters

During the designing of the device, the aim of transistor design optimization has typically been to minimize intrinsic gate delay defined as $CV_{dd}/I_{d,sat}$, where C is gate capacitance in inversion, V_{dd} is the supply voltage and $I_{d,sat}$ is the saturation current. The optimal source-drain separation is determined by the tradeoff between short-channel effects (SCEs) and series resistance (R_s). For a fixed-gate work function (Φ_M), the leakage current increases as the source-drain separation decreases, due to increased short-channel effect. Thus, Φ_M is adjusted in order to meet the I_{off} specification: as the source-drain separation decreases, a higher Φ_M is used to compensate for the increased leakage due to increased short channel effect. When L_g is scaled from 25 nm to 13 nm, the optimal source-drain separation is actually larger than L_g [4]. This indicates that it will be necessary to employ an effective channel length that is larger than the physical gate length in the sub-10 nm L_g.

The gate capacitance consists of the channel capacitance and parasitic overlap and sidewall capacitances [5]. As the source and drain regions come closer together, the overlap capacitance between the gate and the source and drain regions increases. As the height of the gate electrode is increased, the parasitic sidewall capacitance increases. Also, if the raised source and drain is employed, the series resistance is reduced; however, there is an additional contribution to the fringing sidewall capacitance. While a shorter gate height and source and drain regions are desirable for achieving lower sidewall capacitance, their heights are usually determined by sheet resistance requirements in order to keep parasitic resistances low. If the gate height and raised source and drain regions are not scaled with L_g, parasitic capacitances will cause further relative performance deterioration.

As the source separation increases, the parasitic overlap capacitances become smaller and hence the delay decreases. Optimal source-drain separation for minimum delay is larger than that determined for maximum transistor drive current. This is because the effect of reducing parasitic

capacitance is more significant than that of reducing $I_{d,sat}$. When the source-drain separation increases beyond $1.25 \times L_g$, series resistance limits performance, causing the delay to increase with source-drain separation. When a raised source-drain structure is introduced, the parasitic series resistance is reduced, resulting in an optimal source-drain separation that would be even higher than that without raised source-drain. The optimal source-drain separation corresponding to minimal delay should provide for lower dynamic power consumption, because the parasitic portion of the total switching capacitance is lowered significantly.

Thus, the effect of parasitic capacitance on circuit performance is significant, particularly if the thicknesses of the gate electrode and source-drain contact regions are not scaled with the gate length. Therefore, the optimal gate-to-source-drain overlap for maximizing circuit performance will be lesser than that needed to maximize drive current.

Table I: *Parameters Used for Transistor Simulations. These are Essentially Taken from the ITRS (2001 Edn.) except that More Conservative Values of T_{ox} and I_{off} Are Used* [6].

Lg (nm)	25	13
T_{ox} (Å)	11	8
T_{Si} (nm)	7	5
V_{dd}(V)	0.7	0.5
Gate height (nm)	37.5	19.5
S-D gradient (nm/dec)	2.8	1.4
S-D doping (cm^{-3})	$2x10^{20}$	$2x10^{20}$
$\tau_{relaxation}$ (ps) [9]	1.6	1.3
I_{off} (μA/μm)	0.3	1

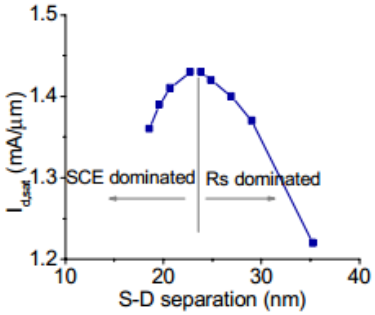

Fig. 12: *Dependence of $I_{d,\ sat\ on}$ Source-to-Drain Separation, for $L_g = 25$ nm. The Separation Is Defined at the Positions where the S/D Dopant Concentration Falls to $2 \times 10^{19}\ cm^{-3}$.The Gate Overlap is Symmetric for Source and Drain.* [4]

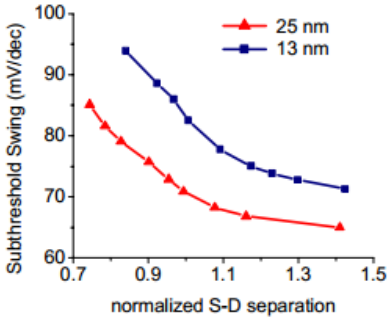

Fig. 13: *Dependence of Sub-threshold Swing on S-D Separation Normalized w. r. t. L_g. As L_g Is Scaled Down, the Relative S-D Separation Required to Control SCE Will Increase* [4]

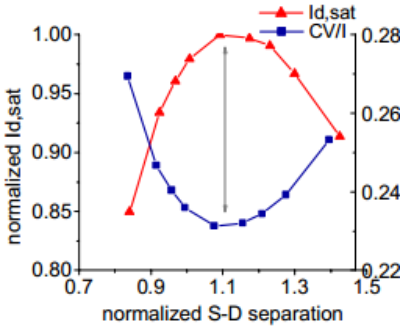

Fig. 14: *Dependence of $I_{d,sat}$ and Intrinsic Gate Delay (CV/I) on the Normalized S-D Separation for $L_g = 25$ nm. $I_{d,sat}$ Is Used henceforth as the Metric in Place of $CV_{dd}/I_{d,sat}$, since Both of Them Point to the Same Optimal Separation. Since C_{max} Remains Constant with Separation, Minimum CV/I ⇒Maximum $I_{d,sat}$* [4].

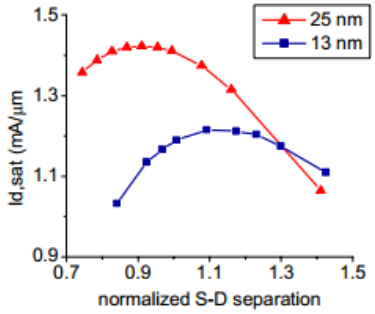

Fig. 15: *Dependence of $I_{d,sat}$ on the Normalized S-D Separation. The Optimal S-D Separation Increases as L_g Is Scaled Down, and Will be Larger than L_g in the Sub-10 nm Regime* [4].

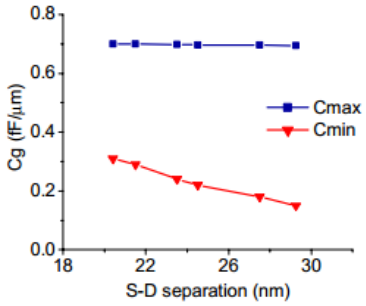

Fig. 16: *Variation of Gate Capacitance (Cg) Parameters vs. S-D Separation for L_g = 25 nm (Assuming a Line Gate, i.e., T_{gate} = 0). While C_{max} (at $V_{gs} = V_{dd}$, V_{ds} = 0) Remains Constant, C_{min}(C_g at $V_{gs} = V_{ds}$ = 0) Decreases Linearly with S-D Separation* [4].

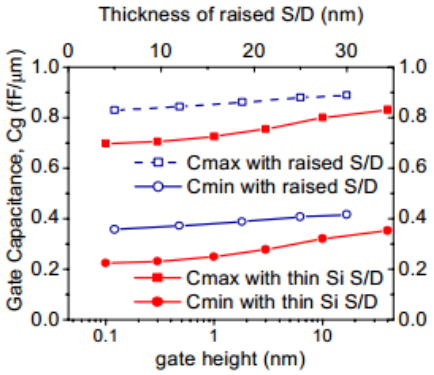

Fig. 17: *Variation of Gate Capacitance (C_g) with the Gate Height (T_{gate}) for L_g = 25 nm and 23.5 nm S-D Separation. In Addition, the Effect of Fringe Capacitance from Raised S/D Regions is Shown as a Function of Thickness of Raised S/D (Shown along the Top X-Axis) for T_{gate} = 37.5 nm* [4].

The double-gate MOSFET and SOI MOSFET have output conductance less than bulk MOSFET. The output conductance and device capacitance can impact device linearity. Output conductance remains relatively flat with the change in drain voltage. When the drain voltage increases, output conductance decreases but in saturation, conductance becomes low but does not reach zero value.

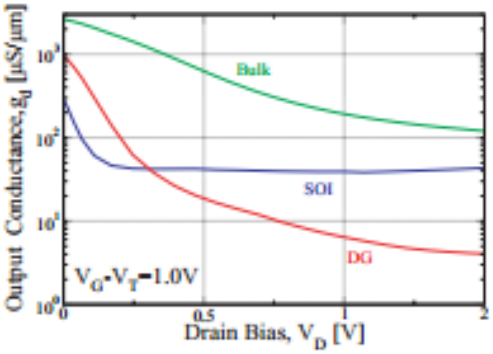

Fig. 18: *Drain Bias Dependence of g_d Is Typically Low in Saturation but Non-Zero. As Device Becomes Thinner g_d Drops* [1].

As the double-gate MOSFET and SOI MOSFET have low thermal conductivity, they face performance degradation because of heat dissipation within the active device. The self-heating effect reduces drain current and produces distortion in the output.

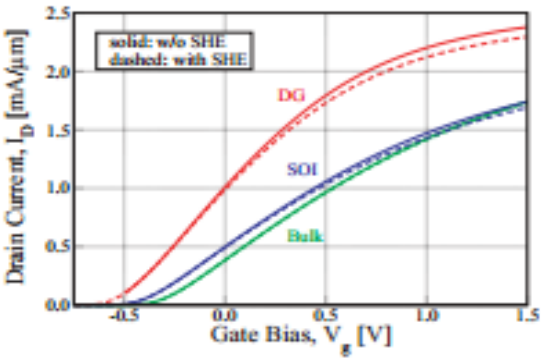

Fig. 19: *I_d-V_g Characteristics with and without Self-Heating Effect* [1].

REFERENCES

1. Fritz J. *Linearity Analysis of Single and Double-Gate Silicon-on-Insulator Metal-Oxide-Semiconductor-Field-Effect-Transistor*. PhD Diss. Ohio University. 2004.
2. Digh Hisamoto, Wen-Chin Lee, Jakub Kedzierski, et al. *IEEE Transactions on Electron Devices.* December 2000. 47p.
3. Chenming Hu, Tsu-Jae King, Vivek Subramanian, et al. *FINFET Transistor Structures Having a Double Gate Channel Extending Vertically from a Substrate And Methods of Manufacture.* U.S. Patent 6,413,802. Jul. 2, 2002.
4. Balasubramanian Sriram, Leland Chang, Borivoje Nikolic. *Proceedings of the 2003 Silicon Nanoelectronics Workshop.* 2003. 16–17p.
5. Rabaey Jan M., Anantha P. Chandrakasan and Borivoje Nikolic. *Digital Integrated Circuits*. Prentice-Hall, 1996.
6. 2001. ITRS. http://public.itrs.net