

ULTRA HIGH SPEED FACTORIAL DESIGN IN SUB-NANOMETER TECHNOLOGY

Arindam Banerjee, Rajan Ghosh, Mainuck Das, Aniruddha Ghosh,
Subham Ghosh, Samayita Sarkar

JIS College of Engineering

ghoshrajan@yahoo.com, mrmmainuckdas@gmail.com, ghoshani19@gmail.com
, subhamghosh86@gmail.com, banerjee.arindam1@gmail.com, samayitaee
@yahoo.com

ABSTRACT

This work proposes a high speed and low power factorial design in 22nm technology and also it counts the effect of sub nano-meter constraints on this circuit. A comparative study for this design has been done for 90nm, 45nm and 22nm technology. The rise in circuit complexity and speed is accompanied by the scaling of MOSFET's. The transistor saturation current I_{dsat} is an important parameter because the transistor current determines the time needed to charge and discharge the capacitive loads on chip, and thus impacts the product speed more than any other transistor parameter. The efficient implementation of a factorial number is carried out by using a decremented and multipliers which has been lucidly discussed in this paper. Normally in a factorial module a number is calculated as the iterative multiplication of the given number to the decremented value of the given number. A Parallel adder based decremented has been proposed for calculating the factorial of any number that also includes 0 and 1. The performances are calculated by using the existing 90-nm CMOS technology and scaling down the existing technology to 45-nm and 22-nm.

KEYWORDS

MOSFET Scaling, Decrementing, Multiplier, Factorial Design.

1. INTRODUCTION

According to the survey of International Technology Roadmap for Semiconductors (ITRS), MOSFET's with 16nm will be available in 2010, 10nm in 2015 and 6nm in 2020 [1]. Up to the 65nm node the optimized scaling strategy (basically equal to the ITRS 2001) is sufficient for the low operating power (LOP) and low standby power (LSP) products to achieve an annual performance increase of 17%-per-year. High performance (HP) products, again, require the most aggressive use of the performance boosters, such as requiring strained-Si channels beginning at the 65-nm node.

The classical MOSFET scaling was first described by Dennard in 1974 [6]. Classical MOSFET scaling techniques were followed successfully until around the 90nm generation, when gate-oxide scaling started to slow down due to increased gate leakage [7]. When gate-oxide thickness can no longer be scaled, then other key of MOSFET parameters, such as supply voltage, cannot be scaled and still expect to deliver improved transistor performance. Without new inventions, MOSFET scaling and Moore's Law were threatened with the likelihood of coming to an end. Intel's 45nm

logic technology was the first to introduce high-K dielectric with metal-gate transistors for improved performance and reduced leakage [8, 9]. Gate-oxide equivalent oxide thickness (EOT) is scaled to 0.9nm and transistor gate pitch is scaled to 112.5nm. Transistor performance is increased by more than 22% compared to 45nm transistors through a combination of increased drive current and reduced gate capacitance [7].

Several architecture of decremented and incremented has been proposed in last few years. S. Bi [3] proposed a multiplexer based incremented and decremented algorithm in 2005, program counter [4], frequency divider [5] etc. The multiplexer based incremented and decremented consist of a serial connection of OR gates. The major drawback of this decremented is it does not include 0 and the series connection of OR gates contribute to a huge delay. For the calculation of factorial '0' and '1', a two stage adder block is used.

Array based multiplier have been proposed by Carboognani in 2005 [3]. Here, parallel adder based multiplier circuit have been used. The main advantage of this multiplier from the previous one is the number of stage reduction. This concept has been adopted and modified for our proposed factorial design. Only a single 4 bit multiplier is used for computing the factorial number.

The design preliminaries of factorial design are consisting 6 sections, MOSFET scaling are introduced in section II, Decrement design are described in section III, Multiplier design are introduced in section IV, section V consisting Factorial design algorithm and simulation results are described in section VI.

1.1 MOSFET SCALING

The goal and challenge of MOSFET scaling is clear: IC density and speed will continue to increase because of inter corporation competition, customer demand, and as a test for the limit of human ingenuity. In the development of an MOS technology, the scaling parameters of the MOSFET's are critical. Such parameters include the channel length (L_c) gate oxide thickness (t_{ox}), channel doping (N), and junction depth (X_j). A generalized design guide had been established as [2]

$$L_{\min} = A \cdot [X_j t_{ox} (w_s + w_d)^2]^{1/3} \quad (1)$$

L_{\min} is the minimum channel length, for which long channel sub-threshold behaviour will be observed. A is the proportionality factor, X_j is the junction depth, t_{ox} is the oxide thickness, w_s is the source depletion depth w_d is the drain depletion depth in a one dimensional abrupt junction formulation. That is $w_d = \sqrt{2} L_B [B(V_{DS} + V_{bi} + V_{BS})]^{1/2}$ where L_B is the bulk Debye length given is by

$L_B = \left[\frac{\epsilon_s}{BqN_A} \right]^{1/2}$. Where $B = (kT/q)^{-1}$, V_{DS} is the drain to source voltage, V_{bi} built in voltage of the junctions and V_{BS} body to source reverse bias. The key parameter of the MOSFET scaling are described Taur, Buchanan in [10]. The issues are: power supply and threshold voltage, short-channel effect, gate oxide, current carrying capability, leakage current limit and interconnect delays.

1.1.1. Power Supply and Threshold Voltage:

In general the active power of a CMOS chip is given by (excluding crossover current) $P_{ac} = (C_{sw} V_{dd}^2 / 2) f$. Where C_{sw} is total node capacitance being switched (either up or down) in a clock cycle and f is the clock frequency. It is clear that by reducing the power supply voltage is dramatic reduction in power dissipation per device.

In contrast of power supply voltage the threshold voltage has not been scaled so much. This is because of standby power requirement. The worst case standby power of a CMOS chip is given by

$$P_{off} = W_{tot} V_{dd} I_{off} = W_{tot} V_{dd} I_0 e^{\left(-\frac{qV_{t,wc}}{mkT} \right)} \quad (2)$$

Where W_{tot} is the total turned off device width, with V_{dd} across it, I_{off} is the off current per device width at 100°C . m is the dimensionless factor normally value ≈ 1.3 , $V_{t,wc}$ worst case threshold voltage at 100°C . The primary reason V_t cannot be scaled because the inverse sub-threshold slope.

1.1.2. Short Channel Effect:

Short-channel effect is the decrease of threshold voltage in short-channel devices due to two-dimensional (2-D) electrostatic charge sharing between the gate and the source drain regions. The short-channel effect plays a key role in V_t -tolerances which determine the minimum acceptable V_t . To scale down MOSFET channel length without excessive short-channel effect, both the oxide thickness and the gate-controlled depletion width (W_{dm}) in silicon must be reduced in proportion to length (L). For the same gate depletion width (W_{dm}), the surface electric field (ξ_x) and the total depletion charge of an extreme retrograde channel is one-half of that of a uniformly doped channel. This reduces threshold voltage and improves mobility. Retrograde channel doping represents a vertically non uniform profile that allows the threshold voltage to be decoupled from the gate-controlled depletion width. However, the body-effect coefficient,

$$m = 1 + \left(\frac{\xi_{Si}}{W_{dm}} \right) \approx 1 + \left(3 \frac{t_{ox}}{W_{dm}} \right) \quad (3)$$

And the inverse sub-threshold slope, $\log_{10} \left(\frac{mkT}{q} \right)$, are still coupled to the gate depletion width W_{dm} . For a given t_{ox} reduction in W_{dm} improves short-channel effect but compromises substrate sensitivity and sub threshold slope. Halo doping or non uniform channel profile in the lateral direction provides yet another degree of freedom which can be tailored to further minimize V_t -tolerances due to short-channel effect [11].

1.1.3. Gate Oxide:

To keep short-channel effect into a desirable limit and to maintain a good sub-threshold turn-off slope, gate oxide thickness is reduced nearly in proportion to channel length [11] as shown in Fig. 1. For $0.1 \mu\text{m}$ CMOS devices operating at 1.5 V , an oxide thickness of about 30 \AA is needed. With such a thin oxide, quantum mechanical (QM) tunnelling takes place, leading to a gate leakage current that increases exponentially as the oxide thickness is scaled down.

To find the scaling limitation of gate oxide, tunnelling current characteristics of ultra-thin oxides have been studied for both measured and simulated current-voltage ($I_g - V_g$) characteristics. Due

to the confinement of their motion in the direction normal to the surface, electrons in the inversion layer must be treated quantum mechanically as a two-dimensional electron gas (2-DEG) rather than classically as a three-dimensional (3-D) gas [12]. The gate leakage current increases exponentially with decreasing oxide thickness. Another issue with the thin gate oxide is the loss of inversion charge, and therefore trans-conductance, due to inversion layer quantization and poly-silicon-gate depletion effects.

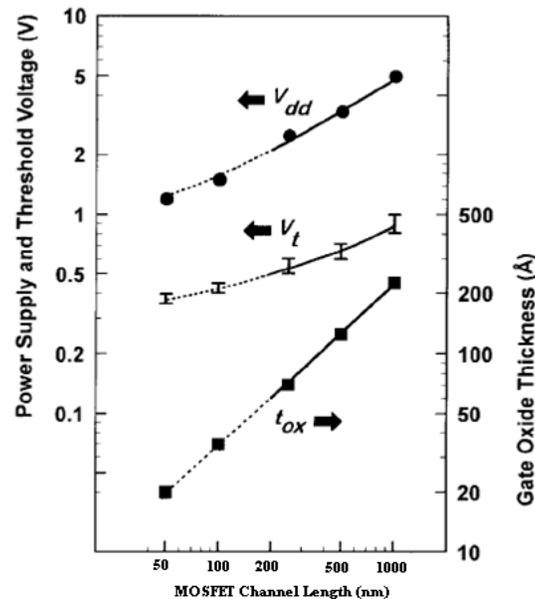


Fig. 1 Power Supply Voltage (V_{dd}), threshold voltage (V_t) and gate oxide thickness (t_{ox}) trends vs. length for cmos technologies.

1.1.4. Current Carrying Capability:

The MOSFET saturation current I_{dsat} can be approximated as

$I_{dsat} = W \cdot V_{sat}$ where \bullet denote the inversion charge density at “pinch off” point.

$$I_{dsat} = W V_{sat} C_{OX} (V_g - V_t - V_{dsat}) \quad (4)$$

$$V_{dsat} = \frac{(V_g - V_t) E_{sat} L_{eff}}{V_g - V_t + E_{sat} L_{eff}} \quad (5)$$

Where W is the per-channel width, V_{dsat} is the drain saturation voltage, V_{sat} is the saturation velocity and E_{sat} is the velocity saturation field. It can be seen from equation (4) and (5) if length approaches to zero I_d holds a constant value. In contrast reducing the oxide thickness, i.e., increasing the oxide field yield a considerable increase in I_d [13].

1.1.5. Interconnect Delays:

If the performance of processors is to keep pace with the speed improvement of devices, special attention has to be paid to interconnections. This is because the delay stemming from wire resistance, commonly referred to as RC delay, to first order is given in eqn. no (6) and has to included with other delay components.8

$$\Gamma = RL_w(C_L + \frac{1}{2}CL_w) \quad (6)$$

Here R and C are the resistance and capacitance of unit wire length and L_w is the wire length and C_L is the load at the end of the wire. The part in the RC delay due to wire alone does not decrease in spite of scaling to smaller dimension. The factor that improves $\frac{RCL_w^2}{2}$ the term through shorter L_w is negated by the increase in R due to wire cross-sectional shrinkage. Wire capacitance in the mean time remains constant, and its value is around 0.2 pF/mm for minimum width wires with oxide dielectric.

1.2 Decrementer

This section proposes a fast adder based Decrementer. Block level implementation of the Decrementer are shown in Fig. 2.

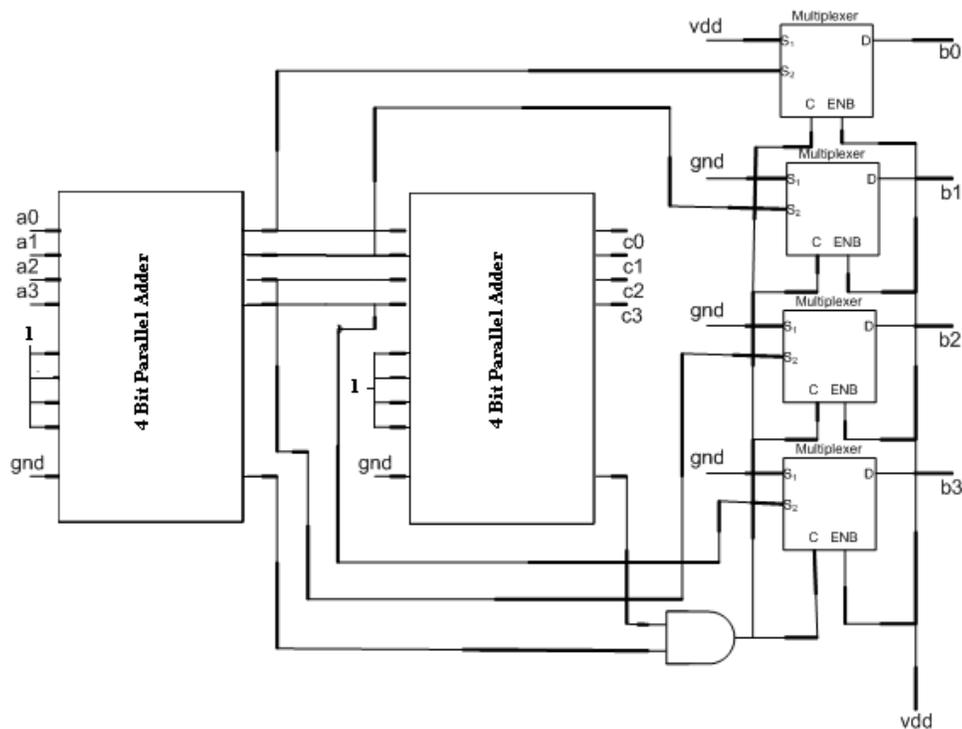


Fig. 2 Architecture of the proposed Decrementer

1.2.1. Decrementer Algorithm

Consider n number of bits and the input of Decrementer is $X = x_{n-1}x_{n-2} \dots x_1x_0$. The formulation of the decremented value can be described as

Begin

Add the data bits to the maximum value of the n bits.

Add the results with maximum value of n bits

If

Both the carry bit is 1 then

Output= first stage addition results.

Else

Output= '1'

End

Example: Consider the input of the Decrementer X= 0101.

For simplifying the calculation the number is taken as 4 bit array, higher bit array can be calculated from the same manner.

The result of the first adder is 10100, here the carry bit is '1', and again added with 1111, and the output of the second adder is 10011, also the carry bit is '1'. Thus from the algorithm the output of the Decrementer is 0100.

The first adder output (without carry) is fed to the second adder stage for the calculation of the factorial number of '1', because the decremented value of the '1' is '0'. Multiplication of both the terms '1' and '0' is always '0' not satisfying the factorial results. The simulation results of the Decrementer are shown in Table 1.

Table 1 Power delay analysis of Decrementer

	90 nm	45 nm	22 nm
Delay (ps)	120	36.6	23.2
Avg. Power (μ W)	5.15	10.2	17.2
Lkg. Power	3.06	14.6	11.6

1.3 Multiplier

Assume the product of two N-bit words are described as

$$x = \sum_{i=0}^{N-1} x_i 2^i \quad (7)$$

$$\text{And } y = \sum_{j=0}^{N-1} y_j 2^j \quad (8)$$

Where $x_i, y_j \in \{0, 1\}$

$$\text{Multiplication can be described as } z = xy = \sum_{i=0}^{N-1} x_i 2^i \sum_{j=0}^{N-1} y_j 2^j = \sum_i \sum_j x_i y_j 2^{i+j} \quad (9)$$

Let $k = i+j$

$$z = \sum_{k=0}^{2N-2} \sum_{i=0}^{N-1} x_i y_{k-i} 2^k \quad (10)$$

$$z = \sum_{k=0}^{2N-2} z_k 2^k \quad (11)$$

$$\text{Where } z_k = \sum_{i=0}^{N-1} x_i y_{k-i} \quad (12)$$

The equation (11) shows that the co-efficient of multiplication can be achieved by the convolution sum of the two finite duration sequences. Considering the long-hand sequences of equation (10) a 4 bit multiplier algorithm are described in Fig. 3. Higher bit multiplier can be described in similar manner. Partial products are added in two stages. Adders and 4 to 3 compressors are used to minimize the stage operations. Compressors and adders are used carefully so that a minimum number of outputs would be generated. Thus by using minimum


```

Add the n bit number to the maximum value of n bit
Add the adder result to the maximum value of n bit
  If (both the carry = '1') then
    Y ≠ '0'
    Decrement Y by '1'
    Result = Y(Y-1).....1
  else
    Add '1'
    Result = '1'
  end
end

```

The hardware implementations of the factorial consist three parts. (i) Zero detectors (ii) Decrementer (iii) Multiplier. Zero detectors are described in Fig. 5. Decrementer and its algorithm were described in section III, and multiplier algorithms were described in section IV.

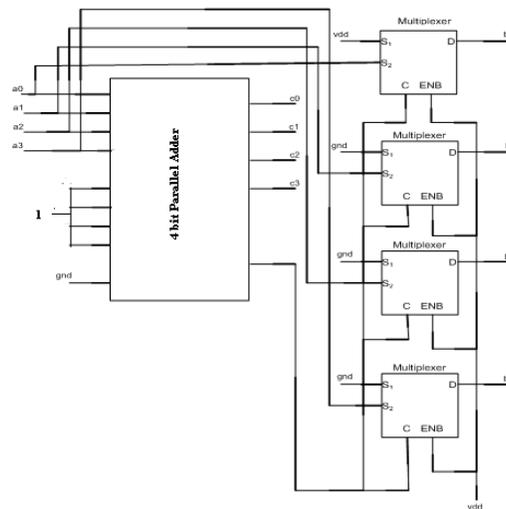


Fig. 5 Hardware architecture for Zero Detector

2. RESULT, ANALYSIS AND DISCUSSION

When a 4 bit number and its decremented result are multiplied the multiplier block gives 8 bit output. The decremented result is then again decremented and multiplied with the 8 bit output. Hence again 8×4 bit multiplier required. This problem has been resolved by recursive multiplication process. The output of the 8×4 bit multiplier is 12 bit, and it is again multiplied by decremented value of the input number. Fig. 4 shows the recursive multiplication architecture of 8×4 bit multiplier. Fig. 6 shows the functional block diagram of a factorial design.

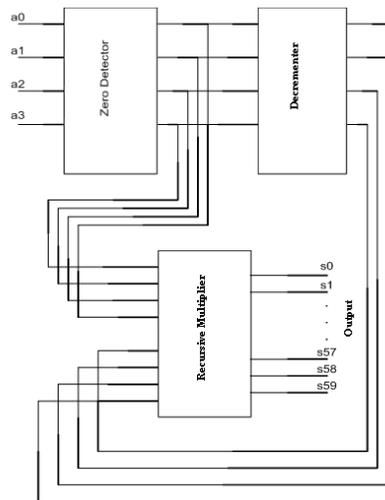


Fig. 6 Functional Block Diagram of Factorial Design

Consider the input number is '0' (decimal value) and its binary representation in 4 bit format is '0000', and passing through the hardware of the factorial design, then starting from the beginning it passes through the zero detector. The output of the zero detectors is (in 4 bit format) '0001'. The special type are assumed here for calculating the values of factorial '0' and '1', because all the mentioned values are '1'. The output of the zero detector is promoted to the input of the decremented and the decremented output value for this special case is also assumed as '0001' (in 4 bit format), because the multiplied value of any number with '0' is always '0'. Then it is multiplied to recursive multiplier block of (4×4) bit, for generating the output values.

Consider again a second block of nonzero input value '5' (decimal representation) and its binary representation are '0101'. First it is passing through the zero detector, output value of the zero detector are same (i.e. '0101') as input number. Then the output of the zero detectors is promoted to the decremented. After first pass of decremented the decremented value is '0100' and it is multiplied to the output of the zero detectors and produced an 8 bit output which is (00010100). For second pass again the value of the decremented is '0011' and it is multiplied again to the output of the first pass. The value from the second stage multiplication is (000000111100). The process will have to be continued until decremented output become '0'. Performance analysis of the factorial design is shown in fig 7.

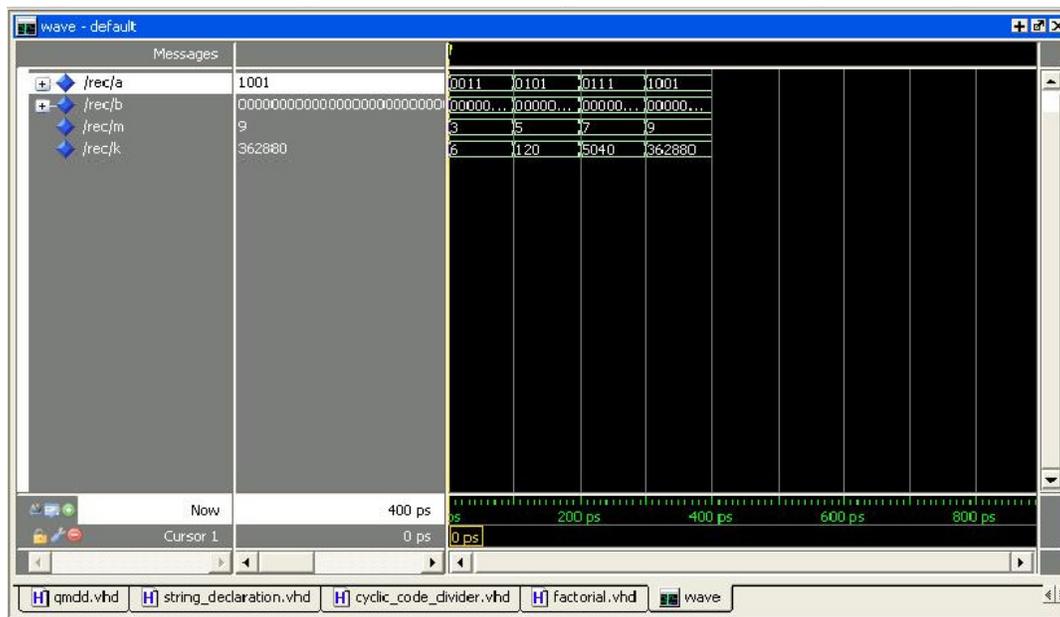


Fig. 7: Performance analysis of the factorial design

3. CONCLUSIONS

By the use of decremented and multiplier we have efficiently implemented the factorial number, the iterative multiplication and parallel adder based decremented are vividly used to generate factorial of any number including 0 and 1.

REFERENCES

- [1] Semiconductor Industry Association (SIA), International Roadmap for Semiconductors, 2003 ed., Austin, TX. Int. SEMATECH, 2003. [Online]. Available: <http://public.itrs.net>
- [2] J. R. Brews, W. Fichtner, E. H. Nicollian, and S. M. Sze, "Generalized guide for MOSFET miniaturization," IEEE Electron Dev. Lett., vol. EDL-1, p. 2, 1980.
- [3] Shaoqiang Bi, Warren J. Gross, Wei Wang, Asim Al-Khalili and M. N. S. Swamy, "An Area-reduced Scheme for Modulo 2^n-1 Addition/Subtraction", Proceedings of the 9th International Database Engineering & Application Symposium (IDEAS'05).
- [4] Carbognani F, Buergin F, Felber N, Kaeslin H, Fichtnes W, "A 2.7/SPL μ W/MHz Transmission-Gate-Based 16-Bit Multiplier for Digital Hearing Aids," in 48th Midwest Symposium on Circuit and Systems, 2005, Zurich, Switzerland, Vol. 2, pp. 1406-1409, Aug 2005.
- [5] Stan M.R, "Long and fast Up/Down counters", IEEE Tran. on Comp., July 98, pp722-735
- [6] Lutz D.R, "Programmable modulo-K counters", IEEE Tran. On Circuits and Systems: Fundamental Theory and Appl., Nov.96, pp.939-941.
- [7] R. Dennard, F. Gaensslen, H. Yu, V. Rideout, A. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," IEEE J. Solid-State Circuits, vol. 9, no. 5, pp. 256-268, Oct. 1974.
- [8] Mark Bohr, "The New Era of Scaling in an SoC World", ISSCC 2009.
- [9] K. Mistry, et al., "A 45nm Logic Technology with High- κ + Metal-Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," IEDM Tech. Dig., pp. 247-250, Dec. 2007.
- [10] C. Auth, et al., "45nm High- κ + Metal gate Strain-Enhanced Transistors," Symp. VLSI Technology, pp. 128-129, June 2008.

- [11] Yuan Taur, Douglas A. Buchanan, Wei Chen, David J. Frank, Khalid E. Ismail, Shih-Hsien Lo, George A. Sai-Halasz, Raman G. Viswanathan, Hsing-Jen C. Wann, Shalom J. Wind And Hon-Sum Wong, "CMOS Scaling into the Nanometer Regime", proceedings of the ieeee, vol. 85, no. 4, april 1997.
- [12] Y. Taur, S. Wind, Y. Mii, Y. Lii, D. Moy, K. Jenkins, C. L.Chen, P. J. Coane, D. Klaus, J. Bucchignano, M. Rosenfield, M. T homson, and M. Polcari, "High performance 0.1 m CMOS devices with 1.5 V power supply," in 1993 IEDM Tech. Dig., pp. 127–130.
- [13] F. Stern, "Self-consistent results for -type Si inversion layers," Phys. Rev. B, vol. 5, No. 12, pp. 4891–4899, 1972.
- [14] Chenming Hu," Future CMOS Scaling and Reliability" Proceedings of IEEE, VOL 81, No 5, 1993

AUTHORS

A. Banerjee finished his M.Tech from WBUT, WB, India in 2008. He is presently working in JIS College of Engineering as an Asst. Professor since 2008. His area of specialization is Microelectronics and VLSI design. His areas of interest are Signal processing, Image processing and embedded system design.



M.Das finished his M.Tech from WBUT, WB, India in 2011. He is presently working in JIS College of Engineering as an Asst. Professor since 2011. His area of specialization is Communication Engineering. His areas of interest are Signal processing and Digital Communication.



A. Ghosh finished his M.Tech from CU, WB, India in 2011. He is presently working in JIS College of Engineering as an Asst. Professor since 2011. His area of specialization is Radio-physics & Electronics. His areas of interest are Signal processing and Digital Communication.



R. Ghosh finished his M.Tech from WBUT, WB, India in 2011. He is presently working in JIS College of Engineering as an Asst. Professor since 2011. His area of specialization is Embedded System. His areas of interests are Embedded System and Devices.



S. Ghosh finished his M.Tech from WBUT, WB, India in 2011. He is presently working in JIS College of Engineering as an Asst. Professor since 2011. His area of specialization is Mobile Communication. His areas of interest s are Digital Communication and OpticalCommunication.



S.Sarkar finished her M.Tech from WBUT, WB, India in 2011. She is presently working in JIS College of Engineering as an Asst. Professor since 2011. Her areas of specialization are Microelectronics and VLSI. Her area of interest is VLSI.

