

A Review on Random Dopant Fluctuation Impact on Within-Die Variation

Athira S¹, Dhanya V², Sreeja S³, Chandni C S⁴

Department of Electronics and Communication Engineering,
Prime College of Engineering, Palakkad, Kerala, India

Abstract— Process variation creates core-speed discrepancy among the core in many-core platforms. Random variation is one of the important components that contribute into core-speed discrepancy. In the paper, a novel technique is proposed that uses footer transistors to reduce the delay and power in a many-core platform. Process variation is due to many fundamental deficiencies, impurities, and imperfections during the fabrication process at the nano-scale technologies. The results of this variation have a direct impact on two key parameters of the CMOS transistor: threshold voltage and gate length, which have major implication on the core speed and power. The random component of this variation is mostly attributed to the random-dopant fluctuation, which results in threshold voltage discrepancy among the cores. The proposed technique reduces the random dopant fluctuation by lowering the dopant density and then compensating the threshold voltage using a footer transistor minimizing the static power dissipation.

Keywords—Process Variation, Random Dopant Fluctuation, Footer transistor, Threshold Voltage, systematic Variation, Random Variation.

I. INTRODUCTION

Process variation is the naturally occurring variation in the attributes of transistors when integrated circuits are fabricated. It becomes particularly important at smaller process nodes (<65 nm) as the variation becomes a larger percentage of the full length or width of the device and as feature sizes approach the fundamental dimensions such as the size of atoms and the wavelength of usable light for patterning lithography masks.

Process-induced variations arise from the imperfection in silicon fabrication, and vary from foundries to foundries. Process variation falls into two categories: die-to-die and within-die. The first category is a variation between different dies or chips. The second category, the focus of this work, is the variation within a single chip. It is also called on-chip or intra-die variation. On-chip, or within-die (WID), process variation can further be classified into two components: random and systematic. The behavior of systematic variation is primarily because of physical parameter variations such as variation due to optical proximity in lithographic process. Non-systematic or random process variation arises from the random nature of the ion-implantation during the fabrication process. Random-dopant fluctuation (RDF) is considered one of the main contributors to the random variation component. Unlike the systematic component, the random component causes variation in the threshold voltage even within neighboring cores. Although, this variation has direct implications on all CMOS device parameters, its impact is usually quantified through two main parameters only, namely the gate length (L) and the threshold voltage (V_t). In a many-

core architecture, the variability of these two parameters results in considerable uncertainty of two vital design constraints: the switching speed and the power consumed by each core. For instance, when a chip experiences this kind of variation, some cores, within the chip, may be fast due to a lower threshold voltage but they are leaky and consume more static power. Other cores may be slow due to higher threshold voltage, but consume less static power. Faced with such variations, a designer may decide to run the entire chip at the speed of the slowest core. As speed and power variations increase due to aggressive scaling, running the entire chip according to the slowest core becomes prohibitive due to major speed and/or power degradation.

Systematic variations are deterministic in nature and are caused by the structure of a particular gate and its topological environment. The systematic variations are the component of variation that can be attributed to a layout or manufacturing equipment related effects. They generally show spatial correlation behavior. Systematic WID variations exhibit high degrees of spatial correlation. Random or non-systematic variations are unpredictable in nature and include random variations in the device length, discrete doping fluctuations and oxide thickness variations. Random variations cannot be attributed to a specific repeatable governing principle. The radius of this variation is comparable to the sizes of individual devices, so each device can vary independently. Random variations are small changes from transistor to transistor typically modeled with a normal distribution.

Random variations stem primarily from two main sources. Non-uniform dopant implantation in the channel depletion region affects threshold voltage, and imperfect control of the lithographic process result in non-deterministic gate lengths. The variations can be summarized as the following categories as in Fig.1.

- Wafer-to-wafer variation is caused, for example, by some change in machine conditions along time of manufacturing apparatus.
- Wafer level variation can be caused by any on-wafer non-uniformity in e.g. temperature and gas flow. Time dependence of lithography exposures may be also responsible.

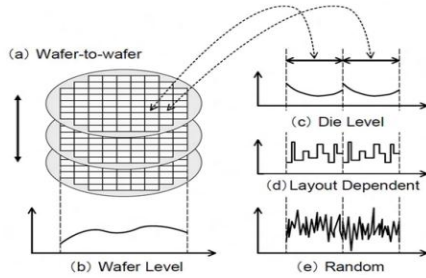


Fig. 1. Classification of variation

- Die level variation typically originates from lithography steps, because pattern exposure is performed die-by-die. It may be caused either by imperfection in reticles or non-ideality in lens systems
- Layout dependent variation exhibits spatial periodicity, as does die level variation, but is different in that it is strongly correlated with the specific layout of patterns, such as density, distance from the neighboring patterns, etc. Pattern dependence of mechanical stress and annealing temperature may be also responsible.
- Random fluctuation is the kind of variability that exhibits no spatial correlation, as already explained.

II. RANDOM DOPANT FLUCTUATIONS

The variations in the structure of a scaled transistor are illustrated in Fig.2. These variations interact with each other, profoundly impacting all aspects of circuit performance. RDF is mainly a random effect. This well known effect is caused by the uncertainty in charge location and numbers, such as the discrete placement of dopant atoms in the channel region that follow a Poisson distribution. As the device size scales down, the total number of channel dopants decreases, resulting in a larger variation of dopant numbers, and significantly impacting threshold voltage.

The channel region of a transistor is doped with impurity atoms. These atoms are randomly placed into the channel by techniques like dopant implantation leading to statistical variations in the actual number of implanted impurities. Such a change of the carrier concentration shifts the threshold voltage and thus the drive strength of the transistor. In older technologies, with thousands of dopant atoms per channel region, an absolute deviation by several atoms was negligible. In recent technologies however the nominal number of impurities is only in the range of tens leading to increased mismatch due to random dopant fluctuation (RDF). Besides the random positioning, fluctuations will occur also in the actual number of dopant atoms present in the channel region. While slight variations on this number are not crucial in sufficiently large channel volumes, they will become critical in deca-nanometre devices showing a moderate doping concentration.

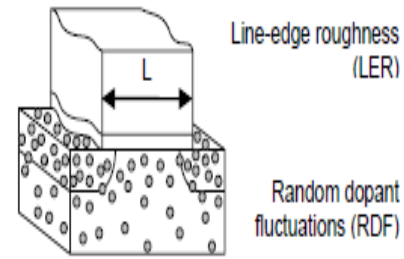


Fig. 2. Primary variation sources in a nanoscale device

For a given technology, RDF is proportional to the dopant density, in the channel region of the CMOS transistor, thus reducing RDF can be realized by reducing the dopant density. Ideally, reducing the dopants to zero would eliminate the RDF completely; however, the CMOS device would no longer function as a controlled switching device, since threshold voltage will be zero. Furthermore, reducing the dopant density decreases the threshold voltage value, which increases the static power consumed by the device. The growing number and complexity of variability mechanisms increase the importance of methods for on-chip measurement. Both systematic and random process variations need to be measured and categorized through silicon measurements. Variability characterization requires collection of a very large amount of data which demands for test structures that are inexpensive in terms of area and test time. Thus, for a given technology, the CMOS device has to be optimized to operate at the lowest possible dopant density, hence lowest RDF value, and then a footer transistor is used to increase the threshold voltage back to the desired value to minimize the static power.

III. PROPOSED TECHNIQUE

The proposed technique reduces the random dopant fluctuation by lowering the dopant density and then compensating the threshold voltage using a footer transistor. The relationship between RDF and the threshold voltage is shown in Fig .3. A very important observation is that, for a given technology RDF is increasing dramatically as threshold voltage increases and especially in smaller technologies (for 9nm, more than 50% increase in σ_{RDF} between =50 and 200mV) as in [1].

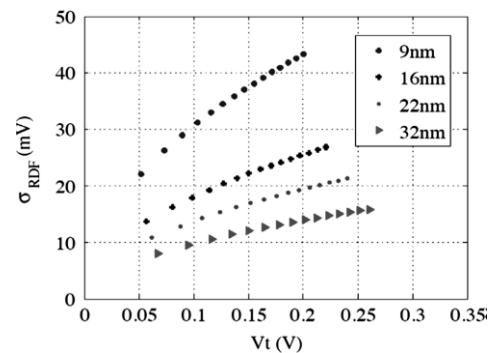


Fig. 3. Graphical representation of RDF variation versus threshold voltage.

Thus, if σ_{RDF} needs to be reduced, the threshold voltage has to be reduced as well. Multi- design, on the other hand, is used normally to increase the threshold voltage in the non-critical logic, which increases the delay but reduces the static power. The threshold voltage, in this case, is determined based on the available slack. In this work, the threshold voltage for each core, in a many- core platform, is reduced lower than the nominal value. Reducing the threshold voltage this way decreases the random variation due to RDF. Then, a footer transistor is used to boost the effective threshold voltage of the core back up to the nominal value. This ensures that the static power is not affected. Since the standard deviation of the RDF is inversely proportional to the transistor size, the footer transistor size is determined such that any RDF incurred by this transistor is reduced.

Normally, the size of the footer transistor should be large, especially in the case of multi-core where each core might have one footer transistor. In the case of large cores, one transistor might be impractical. Instead, each core is clustered into many clusters each with its own optimized footer transistor. This method requires smaller sized footer transistors. In both cases, the footer transistor should be large enough to handle the discharge current of all gates in the cluster. One of the disadvantages of the implementation of footer transistors is the extra area increase. However, as technology scales down, power and speed is considered the main design constraints. In fact, area is considered free compared to power and speed design goals. Thus, using footer transistor is becoming a common practice to reduce power consumption despite the area overhead. However, the proposed method does not attempt to increase the total threshold voltage. In fact, the method tries to divide the nominal threshold voltage between the gate and the footer transistor such that the total threshold voltage is still the same. Thus, the performance of the gate remains the same.

IV. ON CHIP VARIATION

On-Chip variation (OCV) delays vary across a single die due to variations in the manufacturing process (P), voltage (due to IR drop), the temperature (due to local hot spots etc). This need is to be modeled by scaling the coefficients. Delays have uncertainty due to the variation of Process (P), Voltage (V), and Temperature (T) across large dies. On-Chip variation allows you to account for the delay variations due to PVT changes across the die, providing more accurate delay estimates.

The model includes both the systematic and random components, and the variation incurred due to the footer transistor. Random variations cannot be attributed to a specific repeatable governing principle. The radius of this variation is comparable to the sizes of individual devices, so each device can vary independently. In a multi-core architecture, the maximum core speed is inversely proportional to its critical path delay, which is a function of the CMOS device's threshold voltage and the gate length. The critical path can be simply modeled as an inverter chain with the appropriate sizing and loading effects taken into account. The inverter delay is a function of the supply voltage, V_{dd} , threshold voltage, critical path's capacitive load, C_o , and the technology dependent factors α and K . The threshold voltage

in (1) is the equal to the threshold voltage of the critical logic plus the threshold voltage of the footer transistor. The value of K in (2) is dependent on the device characteristics and dimensions, where μ is the mobility, C_{ox} is the oxide capacitance, and W and L are the transistor width and gate length, respectively.

$$D = \frac{C_o V_{dd} / 2}{K(V_{dd} - V_t) \alpha} \quad (1)$$

$$K = \mu C_{ox} (W/L) \quad (2)$$

The systematic variation is modeled as a multivariate normal distribution, and random variation is modeled as uncorrelated normal distribution. This type of modeling is simple enough and it is known to correlate well with empirical data from silicon measurements. The model is used to measure the impact of WID process variation between different cores in a multi-core architecture. The total standard deviation in (3) shows in where both random and systematic standard deviations are added. Since these two components, systematic and random, are formed from two different physical phenomena, they are first computed separately and then added together.

$$\sigma_{total} = \sqrt{(\sigma_{SYS}^2 + \sigma_{RAN}^2)} \quad (3)$$

The systematic process variation for two key transistor parameters, namely the V_t and L_{eff} , can be captured using a multivariate normal distribution with a spatial correlation structure. Thus, the impact of process on the core's critical path delay can be measured through the delay equation considering different V_t and L values generated by the model. Divide the chip into small equally-sized and square-shaped regions. Each region is given a normal distribution for and L_{eff} with mean and a standard deviation. The correlation between two different regions on the chip is dependent on the distance between them. If the correlation function gives a value of zero, then they are uncorrelated. As the distance between the chips increases, the correlation value increases and attains a final value of one when the chips are totally correlated.

The random threshold voltage variation is considered an independent Gaussian random distribution, i.e. the mean is equal to zero. The standard deviation, σ_{RDF} , in (4) of the threshold voltage due to random dopant fluctuations (RDF) is proportional to dopant profile and transistor dimensions given by the equation below, where N_a is the effective channel doping, W_d is the channel depletion region width, L and W the channel length and width and t_{ox} the gate oxide thickness. Based on this equation, as N_a decreases, the variation, i.e. σ_{RDF} , has to decrease. Decreasing the doping density means having lower threshold voltage.

$$\sigma_{RDF} = \frac{q t_{ox}}{\epsilon_{ox}} \sqrt{\frac{N_a W_d}{3 L W}} \quad (4)$$

The other parameters are as follows: V_{fb} , and $2\psi_B$, are technology dependent parameters, and ϵ_{si} is the dielectric constant of silicon. The threshold voltage (5), and the dopant density, N_a , relationship is given by equation below

$$V_t = V_{fb} + \psi_B + \frac{t_{ox}}{\epsilon_{ox}} \sqrt{2\epsilon_{Si} 2\psi_B q N_a} \frac{qt_{ox}}{\epsilon_{ox}} \sqrt{\frac{N_a W_d}{3LW}} \quad (5)$$

V. FOOTER TRANSISTOR SIZING

The variation impact on core speed due to the footer transistor is also dependent on the size of this transistor. So the sizing of the footer transistor is a merely important factor. The size of the footer transistor is proportional to the number of gates discharging through this transistor. Consequently, the size of the footer transistor is always much larger than that of the gate sizes. A larger transistor is expected to have less dopant fluctuations. This is due to the fact that the standard deviation (6) is inversely proportional to the transistor width.

$$\sigma_{footer} = \frac{qt_{ox}}{\epsilon_{ox}} \sqrt{\frac{N_a W_d}{3LW_{footer}}} \quad (6)$$

It is not practical to implement a single transistor for an entire large core. Different methods proposed for footer transistor sizing in a multi-core platform estimates the switching gates percentage to calculate peak current expected to pass through the footer transistor. The size of the footer transistor is proportional to the number of gates switching or discharging through the transistor. The current of a single gate is equal to the saturation current discharging through the pull-down NMOS circuit, assuming NMOS is in saturation mode. The parameters in (7) are the technology dependent factor α , C_{ox} is the oxide capacitance, $V_{t-total}$ is the target threshold voltage ($V_{t-total} = V_{t-Logic} + V_{t-Footer}$), and W and L are the transistor width and gate length, respectively.

$$I_{Gate} = \mu C_{ox} \frac{W}{L} (V_{dd} - V_{t-Total})^\alpha \quad (7)$$

There are two scenarios possible here. The first scenario is to assume that all gates within a core are using the same footer transistor. In that case, the size of the footer transistor should be large enough to handle the discharging of the switching gates. The second scenario is to assume that the core is divided into clusters and each cluster has its own footer transistor. Probably, the gates along the critical path should fall into the same cluster as shown in Fig. 4. In this case, the footer transistor should be smaller than that of the first scenario. Consequently, the variation due to RDF in the case of clustered core should be worse than the case of single footer for the entire core.

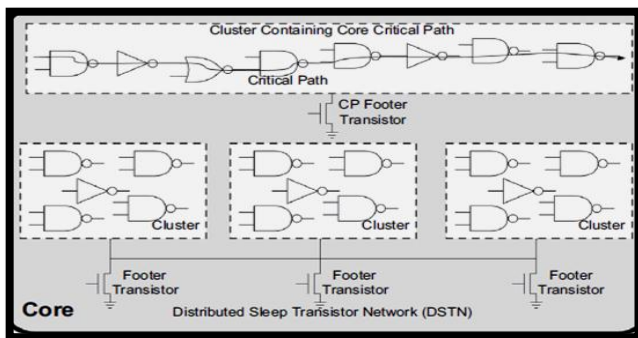


Fig. 4. Traditional clustered core with footer transistors

The proposed method considers the worst-case scenario, i.e. the core is clustered into small regions. Thus, a core is divided into clusters, where each cluster has its optimized footer transistor. The footer transistor is sized based on the

gates in the corresponding cluster. Hence, the size of the footer transistor should be somewhat small. Footer transistors of non-critical clusters are connected together for further area optimization as we discuss later. We assume that all gates along the critical path falls into a single cluster and they all discharge in the same footer transistor. It is focused here on the size of the footer transistor of the cluster containing the gates along the critical path.

$$I_{footer} = \alpha_{switching} N I_{Gate} \quad (8)$$

The total current discharging in the footer transistor is given by (8). The variable $\alpha_{switching}$ is the percentage of switching gates in the corresponding cluster, N is the number of the total gates in the cluster, and I_{Gates} is the discharging current of a single gate. The total discharging current into the sleep transistor is also equal to the current passing through the footer transistor operating in the linear region as in (9).

$$I_{footer} = \mu C_{ox} \frac{W_{footer}}{L} (V_{dd} - V_{t-footer}) V_{t-footer} \quad (9)$$

Since the target delay is still the same, i.e. before and after footer transistor insertion, the source-drain voltage drop of the footer transistor has to be equal to $-V_{t-footer}$. This insures that the gate-delay in the logic stays the same. Note that in traditional footer transistor calculations, the footer's source-drain voltage drop is used to optimize the logic delay. The nominal size of the footer transistor can then be calculated as shown in (10).

$$W_{footer} = \frac{\alpha_{switching} N I_{Gate}}{\mu C_{ox} \frac{1}{L} (V_{dd} - V_{t-footer}) V_{t-footer}} \quad (10)$$

The area overhead due to sleep transistor might be large. Different methods in the literature are proposed to reduce the area overhead. One of the effective methods is using Distributed Sleep Transistor Network (DSTN). In this case, all footer transistors of the non-critical clusters are grouped in a single network. This reduces the size of the footer transistors of these clusters significantly. DSTN is intrinsically better than the cluster-based design in terms of the sleep transistor area and circuit performance. The algorithm obtains DSTN designs with up to 70.7 % sleep transistor area reduction compared to cluster-based designs. The assumption here is that not all the gates in these clusters are discharging at the same time. Thus, footer transistors of the non-critical clusters are transistors with normal sizes. The only footer transistor that should take large area is the one containing the critical path. After calculating the size of the footer transistor, the RDF due to the footer is calculated and a $-V_{t-footer}$ population is generated. The total threshold voltage then is equal to the logic and footer threshold voltage (11).

$$V_{t-Total} = V_{t-logic} + V_{t-footer} \quad (11)$$

After generating the, $V_{t-total}$, population, the critical path delay of each core then is calculated using the delay equation (12). The critical path is the path with longest delay. The threshold voltage of the footer transistor is determined based on the target threshold voltage. For instance, if the target threshold voltage is 400mV, and the core's is reduced to 250 mV, then the V_t of the footer transistor is 150mV.

$$D = \frac{C_0 V_{dd/2}}{k((V_{dd} - V_{t-footer}) - V_{t-logic})^\alpha}$$

$$= \frac{C_0 V_{dd/2}}{k(V_{dd} - V_{t-Total})^\alpha} \quad (12)$$

In a multi-core architectures, WID process variation manifest itself at the granularity of a core. The multi-core architectures proceed the computation in parallel, lessening the frequency requirements for individual cores. Here, performance is achieved by adding cores instead of increasing the clock frequency. Reducing the frequency of individual cores is particularly beneficial from the power point of view. Thus, the model can be extended to a multi-core platform by dividing the chip into separate core regions. A 1024 core with normalized frequency impact due to WID variation is illustrated in Fig .5.

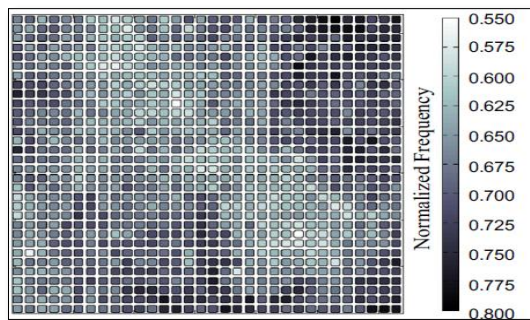


Fig. 5. A 1024 many-core with normalized core frequency due to WID variation

The process variation within each core region has a direct impact on the speed of the core. Also, each core is assumed to have a footer transistor. Based on the variation across a core and its location on the chip, the maximum frequency can be computed for each core. Process variation causes a frequency variation among cores in the same platform. This asymmetric distribution of frequencies creates major scheduling and synchronization problems that can result in system malfunction. WID systematic variation, WID_{sys} , play an important role, because at the 45nm generation and beyond, reduced core areas will cause parameters within a core to be highly spatially correlated, while the amount of variation that can occur across a chip can be large. Systematic variation will result in both C2C frequency and leakage variation. C2C frequency variations will be modest in comparison to leakage variation. This is because the amount of WID_{sys} that occurs across a chip 10–15% variation in gate length has only a linear impact on frequency. Instead, leakage which has an exponential dependence on the gate-length variation (because of its impact on threshold voltage) shows the most important architectural WID variation.

Cores are becoming sufficiently small with technology scaling that spatially correlated phenomena like optical-field variations can introduce significant “systematic” WID variations that produce significant core-to-core (C2C) frequency asymmetry. The choice of floor plan has an important effect on core-to-core asymmetry. When cores are distributed across a large die, they are vulnerable to WID systematic variations. When cores are placed close to each other, the increased power density incurs a greater risk of thermal throttling. This creates a multidimensional tradeoff space among core power, floor plan, magnitude of cross-chip variation, and cooling cost.

VI. PROCESS VARIATION IMPACT

As process technology is moving toward smaller dimensions, the impact of manufacturing defects and variations increases sharply. Multi-core architectures introduce a new granularity at which process variations may occur, yielding asymmetry among cores that were designed. Process variations cause maximum clockable frequency and power dissipation of a high-performance chip to vary from the target frequency and from chip to chip. Post manufacture testing is used to characterize chips and identify the best operating frequency for each. Unfortunately, faster chips usually have higher sub-threshold leakage currents, because the main contributor to frequency variations, L_{eff} , also affects sub-threshold leakage. In fact, the fastest chips often cannot operate at their peak sustainable frequency because the excessive leakage causes the chip to overheat, and a suitable cooling solution may be too expensive. Individual cores are now small enough that the chief impact of many spatially correlated phenomena manifests across rather than within cores. This is a problem because multi-core chips with non-uniform frequency or power characteristics from core to core create scheduling and thermal-management problems. This can cause reduced throughput, missed real-time deadlines, or excessive thermal throttling if more computationally intensive threads are mapped to higher-power cores.

VII. CONCLUSION

A novel technique to reduce the random dopant fluctuation by reducing the threshold voltage, through lowering the dopant density is presented, for a given technology. The findings show the frequency distribution due to process variation, in a many-core, shifts towards the nominal frequency and it is slightly narrowing as the dopant density decreased. In summary, with lower dopant density the impact on the speed variation is reduced. The proposed method in this paper does not attempt to increase the total threshold voltage. In fact, our method tries to divide the nominal threshold voltage between the gate and the footer transistor such that the total threshold voltage is still the same. Thus, the performance of the gate remains the same. It is estimated that the standard variation on core’s frequency variations is reduced and improves the energy saving on a many-core platform.

ACKNOWLEDGMENT

The authors are grateful for the assistance provided by Department of ECE of Prime college of Engineering, Palakkad.

REFERENCES

- [1] A. Agarwal, S.K. Bhunia, J.D. Gallagher, K. Roy, "Effectiveness of low power dual- V_{th} designs in nano-scale technologies under process parameter variations", Proceedings of the International Symposium on Low Power Electronics and Design ISLPED, pp.14–19, 2005.
- [2] Arash Rezapou and Pegah Rezapou, "The Effect of Random Dopant Fluctuation on Threshold Voltage and Drain Current Variation in Junctionless Nanotransistors", Journal of Semiconductors, Vol. 36, No. 9, 2015.
- [3] Abdoul Rjoub, Samer Khasawneh, Mutasem Ajlouni, "Efficient Techniques for Low Power Leakage Current Based on Header/Footer Techniques in Nano-scale Circuits", 9th international Arab Conference on Information Technology, Dec 2009.
- [4] Eric Humenay, David Tarjan, Kevin Skadron, "Impact of Process Variations on Multicore Performance Symmetry", Design, Automation & Test in Europe Conference & Exhibition, IEEE, pp1 – 6, 2007.
- [5] Farshad Moradi, Tuan Vu Cao, ElenaI. Vatajelu, Ali Peiravi, Hamid Mahmoodi , Dag T. Wisland , " Domino logic designs for high-performance and leakage-tolerant applications", INTEGRATION- The VLSI Journal, Vol 46 , pp 247–254, 2003 .
- [6] Jiménez A., Ambrosio R. C., Mireles J. Jr., et al., "Analysis of threshold voltage fluctuations due to short channel and random doping effects", SMCSYV 2013.
- [7] K. G. Verma, "Effect of Process Variation on the Performance Parameter of VLSI Interconnects.", A Final Thesis, School of ECE, SHOBIT University, Meerut, 2011
- [8] Mahroo Zandrahimi, Zaid Al-Ars, "A Survey on Low-Power Techniques for Single and Multi-core Systems' 3rd International Conference on Context-Aware Systems and Applications, March 2015.
- [9] Ming-Hung Han, Yiming Li, et al. , "Comprehensive Examination Of Intrinsic Parameter-Induced Characteristic Fluctuations in 16-nm-gate Devices", NSTI, NANOTECH ISBN, 978-1-4398-3402-2, Vol.2, 201 .
- [10] P. Brahmini, Giri Babu, "Low Power Cmos Design with Sleep Transistor For Submicron VLSI Technologies", International Journal of Eminent engineering Technologies , Vol 1, Issue 4, 2014.
- [11] Sohaib Majzoub, " Reducing random-dopant fluctuation impact using footer transistors in many-core systems", INTEGRATION, The VLSI Journal Vol 48, pp 46–54, 2015.
- [12] Wei Huang, Shougata Ghosh, Siva Velusamy, et al., "HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design", IEEE Transactions On Very Large Scale Integration (VLSI) Systems, VOL. 14, NO. 5, MAY 2006.
- [13] Yun Ye, Frank Liu, Sani Nassif, Yu Cao, "Statistical Modeling and Simulation of Threshold Variation under Dopant Fluctuations and Line-Edge Roughness", Proceedings of the 45th Design Automation Conference, Pages: 900-905, June 2008.