Grenze International Journal of Engineering and Technology, Jan Issue

Chip Variations in Fabrication: Causes, Effects, and Mitigation Techniques for Enhanced Performance

Riya Kapatia[1] and Sateesh Kumar Awasthi[2]
[1-2]Dept. of ECE, Dr B R Ambedkar National Institute of Technology, Jalandhar, India
Email: riyak.ec.21@nitj.ac.in, awasthisk@nitj.ac.in

***Abstract*— In the present-day, the chip industry stands as the economic cornerstone of the modern world. The primary objective of the VLSI semiconductor industry is to attain optimal performance while minimizing costs and power consumption. With rising competition among semiconductor companies, there has been a significant focus on reducing MOSFET size and adopting emerging technology nodes. However, these new nodes may not be fully optimized for high yield in mass production. The need of the hour is to devise a technique that can enhance yield in the semiconductor field, striking a balance between quality and functionality. The quality of results (QoR) for current Application-Specific Integrated Circuit (ASIC) methods has been observed to be significantly influenced by variations. Post physical design and sign-off, the chip is sent for fabrication, where it is susceptible to chip variations, broadly categorized as "global process variation" and "on-chip variation". Neglecting chip variations during the design phase can lead to timing violations and power failures. During this work, chip variations are briefly addressed with possible aids.**

***Index Terms*— Process variation, On Chip Variation (OCV), Complementary Metal Oxide Semiconductor (CMOS) design, Monte Carlo Simulation.**

I. INTRODUCTION

To fulfil the increasing demand of electronics, we largely depend on the semiconductor industry. Chip manufacturing is the driving force of semiconductor market. The increase in demand for power- efficient and high-performance devices within less area, drastically decreases chip size. As we further scale the CMOS device, performance is impacted by second order effects and fabrication becomes difficult. Fabrication process involves the creation of the chip's physical layout and the deposition of various layers and materials. After the completion of physical design and sign-off stages, the chip undergoes the fabrication process, during which its functionality can be influenced by a range of factors. These factors may include process variations, temperature variations, and electrical noise. In the real physical world, the delays associated with cell and wire components are significantly influenced by various factors in the surrounding environment. The cause of variation can be channel length variation, oxide thickness variation, or voltage threshold variation. There are two types of variations: local variation and global variation. It can be inter-die, intra-die, random, or systematic. As shown in Fig 1. The same type of wafers placed near each other can be affected differently.

II. BACKGROUND

The quality of chip fabrication directly impacts the performance, power consumption, and reliability of the chip.

Grenze ID: 01.GIJET.10.1.82

information on timing, power, and IR-drop. This information is used to analyze Power- Performance-Area (PPA) trade-offs and design flows accordingly. Certain variation models are utilized during semiconductor manufacturing to simulate and understand the effects of various variations on chip performance. The On Chip Variation (OCV) model was introduced for nodes above 90 nm. This follows the approach in which additional global delay is applied to all the standard cells in the design regardless of the slew, load and formation of cells. On applying the global derating model, it becomes too pessimistic for longer paths and positive for shorter paths. For lower nodes, OCV model results were inconsistent in providing better performance. The industry then shifted to Advanced on Chip Variation technology (AOCV). In this approach, cell specifications are derated based on distance, cell type, and logical depth. The space of the cell in the path model tends to be a systematic variation of cell while the depth of the cell model is a random variation [6]. This approach becomes incorrect below 40 nm so Parametric on Chip Variation (POCV) is used. POCV uses delay variation of a specific cell to calculate cell delays. It does not add any derates but instead uses the Gaussian curve directly. POCV has reduced pessimism in nodes 20 nm and below and is an integral part of main stream sign-off methodology. The Variations are briefly described and defined in the following sections, along with relevant solutions.

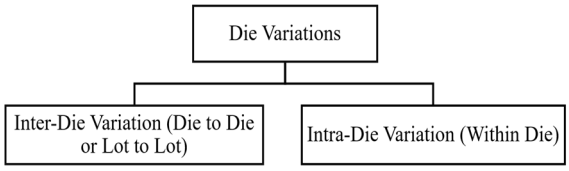

Fig. 1: Die Variations

III. On Chip Variation

On-chip variation model is the process of figuring out how different IC process parameters affect delay, power, leakage, and other things. These can occur mainly due to variations in the manufacturing process, voltage, and temperature as shown in Fig 2. There is a shift in manufacturing conditions due to which one batch of wafers is slow or fast relative to nominal estimates. This is modelled using positive or negative derate numbers. Derate is applied to data clock paths, data paths, and early and late paths for fall and rise transitions during the analysis phase.

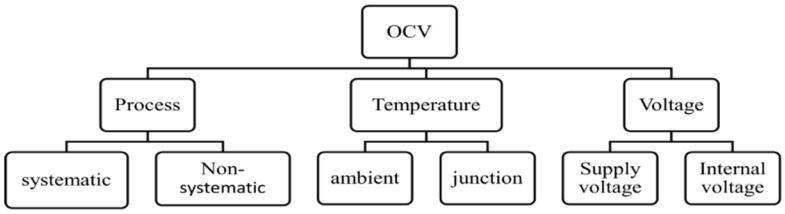

Fig. 2: On Chip Variations

A. Process variation

Process variation is due to deviations in the fabrication process. Deviations vary from wafer to wafer and from one die to another. These variations lead to alterations in the channel length, oxide thickness, and threshold voltage of the CMOS device [1]. They are caused by non-uniformity during deposition or diffusion of impurities. Sources of variation during fabrication manufacturing are generally etching and oxide thickness. Etching is used to define the structure of a transistor. The width and length of the transistor have to be as expected (W and L ratio).

$$I_d = \mu_n C_{ox} * W/L * (V_{GS} - V_{TH}) V_{DS} - V^2_{DS}/2 \quad (1)$$

A small change in W/L leads to a change in drain current, this is shown in 1 [9]. After change in drain current, delay variation is increased, and it impacts the frequency performance of the design.ICs are produced in huge lots. Due to the effect of variations, the electrical properties of different lots can be dissimilar. The same type of CMOS device placed next to each other could have different behavior on the same die. This also impacts turn-around time and yields (the number of chips per wafer that achieve the target frequency).

B. Temperature Variation

For lower nodes, there is a nonlinear situation when we consider temperature fluctuations. The cell propagation gets lower as the temperature goes up. Power dissipation in MOS transistors leads to temperature variations across the chip [8]. The cell delay can be roughly described by 2 [3]

$$Cell\ delay = (C_o * V_{DD}/\mu\ (T)) * (V_{DD} - V_{TH}) \qquad (2)$$

where μ is the electron mobility, and V_{TH} is the threshold voltage [3]. As temperature increases, boththe mobility and Vth decrease. Temperature now depends on the concentration of dopants whenevermobility starts to decrease. Higher doping concentrations result in a high starting temperature. Dueto slower holes and electrons, propagation delay grows with an increase in temperature. Generally, mobility has a greater impact on the overall cell delay. To guard the whole design, we need to know the best and worst performing cases. The chip's best case-scenario has a faster process, greater voltage, and lower temperature. A worst-case scenario has high temperatures, low voltage, and a slow process.

C. Voltage Variation

Internal voltage deviation occurs in the chip based on the design. Generally, the higher the voltage, the faster the cell, this linear relation is used for gate level performance calculations. The power delivery network (PDN), known as the power grid, is made up of metal stripes and rails that transport power to all standard cells inside the chip. The power supply of the CMOS device is non uniform. It is very difficult to keep the distance between the power pad and cells the same for all the standard cells. Depending on the design, there will be a variation in the power supply for the cells. Cell delay depends inversely on available VDD. The amplitude of the voltage drop is given in 3, where L is the self-inductance and I is the current through the line.

$$V=L*\frac{dI}{dt} \qquad (3)$$

IV. Advanced On Chip Variation

OCV is an approach that applies a global blanket margin to guard the design. It has growing concerns such as overdesign, reduced design performance, and longer timing closures, which makeit less acceptable. AOCV tends to provide the right balance of runtime and accuracy for the design[7]. It uses specific derates instead of global derates.

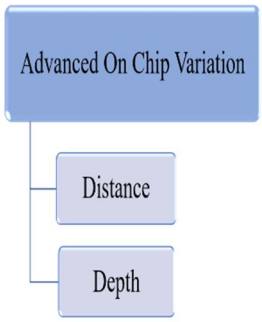

Fig. 3: Advanced on Chip Variations

Here the derate value is based on the net location and logic depth for particular cells [2] as depicted in Fig 3 The logic depth affects random variation. Systematic variation is applicable to the cell where the path under analysis is located [5]. Based on cell depth, advanced OCV derate factors can be determined. This decreases the path's pessimistic margins. The silicon properties affect the derate factors for systematic variation. Random variations generally follow a normal distribution. The modelling of these variations is based onin-depth randomized trials of how variation relates to the geometric separations [10].

V. Parametric On Chip Variation

In POCV, we apply derate based on cell delay variation and using multiple probability simulations. Monte Carlo simulation is used to predict delay outcomes based on variations. The number of standard cells that fall into a certain delay range roughly follows the normal distribution curve. Nominal delay data appears at the

center of the curve, and slow and fast appear at opposite edges. Variations from the normal distribution curve are calculated using the mathematical notion of standard deviation. as in Fig 4.

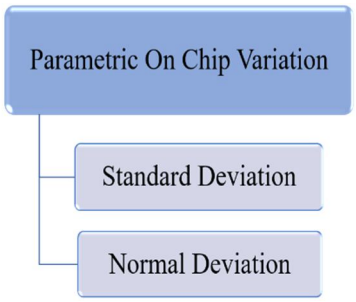

Fig. 4: Parametric on Chip Variations

The standard deviation checks the extent to which the data at the edges or centre isshifted from the normal distribution. Minimum and maximum delay values are used to model random variations. There is a chance that the true delay value will fall outside of the nominal rangeor anywhere between the mix-max extremes. The extreme process corners are defined as the plusor minus three sigma corners from the nominal process variation distribution. Fig 5 shows the sigma concept. Worst case low and best case fast are the two that need the most attention since they are the three sigma extreme corners and guard the whole design. The slow process models correspond to the plus three-sigma corner condition for inter-die variations and fast process models correspond to minus three sigma corner condition for the inter-die variations [4].

VI. Managing Chip Variations

To reduce chip variations in semiconductor fabrication, several proposed methods are employed. Process control measures are implemented to ensure consistency and repeatability during manufacturing, with critical parameters closely monitored and adjusted at each step. Advanced techniques like Monte Carlo simulation statistically analyze the impact of process variations on chip performance, providing valuable insights to optimize designs for robustness against variations. Through the integration of these techniques, chip designers can effectively manage and reduce variations, ensuring higher yields, improved reliability, and enhanced overall chip performance. Library characterization involves characterizing standard cell libraries to predict chip behaviour under different conditions and mitigate the impact of process variations. Design for Manufacturability (DFM) techniques are utilized, optimizing chip layouts to minimize the impact of manufacturing fluctuations and designing features that are less sensitive to variations. Statistical Process Control (SPC) methods help monitor and control variations, identifying trends and enabling timely corrective actions. Additionally, dummy fillers are added to balance stress and strain, reducing variations resulting from different feature densities.

A. Monte Carlo

Monte Carlo simulation is a computational technique used to analyze complex phenomena by generating random samples and observing their statistical outcomes. By simulating a large number of random events or variables, Monte Carlo simulation provides a robust way to estimate probabilities and evaluate the behaviour of a system under uncertainty. This simulation method is widely applied in various fields, including engineering, physics, and risk analysis, to solve problems that involve probabilistic and stochastic elements. Most of the time, Monte Carlo simulations are used to figure out how process variation affects cell delay. Monte Carlo Synthesis approximates the random variation effects usually encountered in Intra-Wafer manufacturing. Maximum Likelihood Estimation (MLE) is used to obtain the gaussian mean μ and standard deviation σ of Normal-Gaussian distribution shown in Fig 5. Using μ and σ values computed by MLE, random timing derate factors are generated and applied per circuit standard cell [2]. The number of standard cells that fall into a certain delay range roughly follows the Gaussian distribution. We tend to be pessimistic when estimating timing verification. To protect our design, we choose the nominal case, the slow case, and the worst corner case. So, for the impact of global process variation, we end up with three corners, which are fast, typical, and slow around the central line of each corner. The variation is observed, but with a much lower

standard deviation. Two steps are followed to make a CMOS device, transistor formation in the base layer and metallization on the higher layer. Due to process variation, a single CMOS device will be either slow or fast, or anywhere in between, across different dies. Thus, transistor formation, where each gate consists of NMOS and PMOS, has four combinations as shown in Fig 6 slowest NMOS and slowest PMOS, slowest NMOS and fastest PMOS, fastest NMOS and slowest PMOS, fastest NMOS and fastest PMOS.

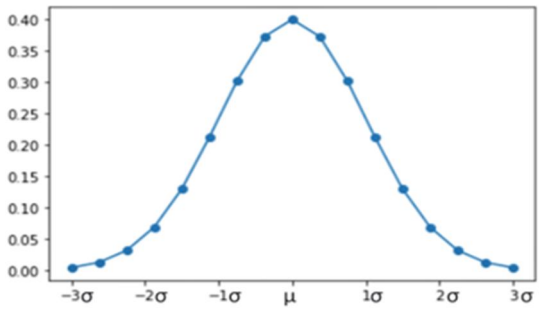

Fig. 5: Gaussian Curve

B. Library Characterisation

Library characterization refers to the process of extracting and representing the essential information about electronic design library components, such as cells, macros, or IPs, to enable accurate and efficient circuit simulation and synthesis. It involves characterizing the behavior and electrical properties of library components across different operating conditions, including power supply voltages, temperatures, and process variations.

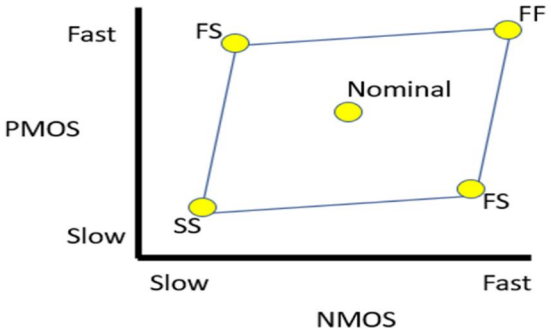

Fig. 6: Corners

Library characterization is typically involved in tasks such as timing analysis, power analysis, noise analysis, and modeling of various electrical parameters, such as propagation delay, setup and hold times, power consumption, and slew rates. The cell analysis tool supports voltage and temperature scaling, which allows for analysis of process, voltage, and timing conditions that are not being characterized. The cell analysis tool uses interpolation to derive appropriate timing, power, and noise values for the current voltage condition. One-dimensional scaling requires two sets of libraries, so that only one characterization parameter, voltage or temperature is varying, and the remaining values are held consistent across all libraries. Two-dimensional scaling needs at least three libraries to accommodate two parameters that vary at the same time. Normally, to achieve good accuracy, there are two libraries per dimension. In total, we need four libraries to sup-port the two-dimensional scaling. Three-dimensional scaling is not used too often due to the fact thatit involves a lot of library characterization work to be done. It allows three independent parametersto be varying at the same time at the cost of much more library sample points to start cell analysis.

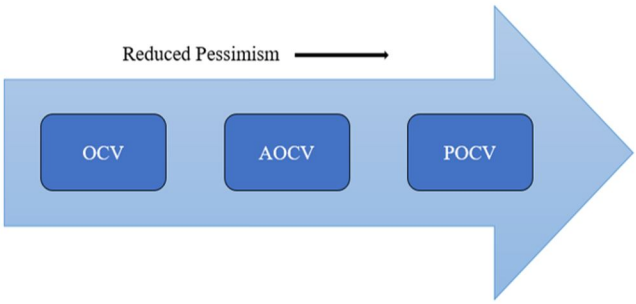

Fig. 7: Pessimism Reduction

VI. Conclusion

The challenges posed by semiconductor processes and manufacturing variability have significantly impacted timing analysis in integrated circuits (ICs). Factors such as temperature variations andminor process changes contribute to variations in IC performance within a wafer batch, deviating from nominal estimates. To address these challenges, OCV modeling has been introduced, enabling the implementation of timing constraints during wafer fabrication and affecting static timing analysis and IR drop considerations. The adoption of AOCV has further improved derating techniques by applying cell and wire-specific derating instead of a single global derate value. By considering the logical depth feature, AOCV acknowledges that cells within deep pathways may exhibit different delays. For advanced technology nodes such as 16/14 nm, POCV has emerged as a more effectiveand efficient approach compared to the overly pessimistic OCV modeling. By integrating POCV with static timing analysis and power-performance-area (PPA) analysis, runtime can be reduced, and the limitations of traditional OCV modeling can be addressed. The combined utilization of POCV and AOCV derating techniques enables the simulation of systematic variances, allowing for more accurateanalysis and optimization of IC designs. The incorporation of AOCV, POCV, and advanced library techniques represents significant advancements in addressing timing challenges caused by process variations. These methodologies enable designers to improve chip performance and reliability by considering systematic variances in ICs and optimizing design margins more effectively.

References

[1] Mahmoud Bennaser, Yao Guo, and Csaba Andras Moritz. “Data memory subsystem resilient to process variations”. In: IEEE transactions on very large scale integration (VLSI) systems 16.12 (2008), pp. 1631–1638.
[2] Nikolaos Blias et al. “Investigation on Performance, Power, Area Trade-Offs using Determin- istic and Monte-Carlo Process Variation Aware Synthesis Flows”. In: 2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC). IEEE. 2022, pp. 1–6.
[3] Andrea Calimera et al. “Temperature-insensitive synthesis using multi-vt libraries”. In: Proceedings of the 18th ACM Great Lakes symposium on VLSI. 2008, pp. 5–10.
[4] Rakesh Chadha and J Bhasker. Static Timing Analysis for Nanometer Designs: A Practical Approach. Springer, 2009.
[5] Rabab Ezz-Eldin, Magdy A El-Moursy, and Hesham FA Hamed. “Analysis and design of Network on Chip under high process variation”. In: 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS). IEEE. 2015, pp. 508–509.
[6] Sajja Krishna Kishore et al. “An On-chip Analysis of the VLSI designs under Process Varia- tions”. In: 2020 International Conference on Smart Electronics and Communication (ICOSEC) (2020), pp. 1273–1277.
[7] Subrat Mahalik. “Automating Variation and Repeater Analysis in Physical Design of Integrated Circuits”. In: (2018).
[8] Rashad Ramzan and Jerzy Dabrowski. “RF calibration of on-chip DfT chain by DC stimuli and statistical multivariate regression technique”. In: Integration 49 (2015), pp. 14–21.
[9] Behzad Razavi. Design of analog CMOS integrated circuits., 2005.
[10] Sunil Walia. “Primetime® advanced ocv technology”. In: Synopsys, Inc (2009).