

Chapter 2

Process Variability

Overview

Parameter variability has always been an issue in integrated circuits. However, comparing with the size of devices, it is relatively increasing with technology evolution, as the device size shrinks in a larger scale than our control over them. Also, in the past, the variations were mostly due to imperfect process control, but now intrinsic atomistic variations become more important, as devices of atomic sizes are achieved. This parameter variation causes uncertainties in circuit design, as in timing, power dissipation, and others important properties. Figure 2.1 shows the technology scaling, to exemplify how small the devices are becoming. Approaching the atomic scale is very difficult to control the process, as only one atom can make a huge difference.

On the other hand, new research from Intel at [25], show that variability is not an insurmountable barrier to technology development, but just a new challenge to overcome and cope with in technology fabrication and design. They successfully showed that it is possible to remain variability under control up to 45 nm, and presented design mitigation techniques. In their experiments, systematic variability was constant at about 3 % of the maximum frequency, and random variability increased from 130 to 90 nm and 65 nm at a maximum of 2.5 %, before decreasing to about 1 % at 45 nm, when they introduced new fabrication control.

This chapter will address the main aspects of process variability. First, we will identify the different sources and types of variations. Then we enter in specific issues that cause variations, as lithography and doping. Finally, we must analyze the true impact on circuit design.

2.1 Sources and Types of Variations

Variations can be classified in many different ways. For example, they can be divided in different sources and types, as in [17]; or between systematic or random sources of variations, as in [39].

The variations can have different sources and types, as presented in [17]. The sources can be divided between process variations, environment variations and

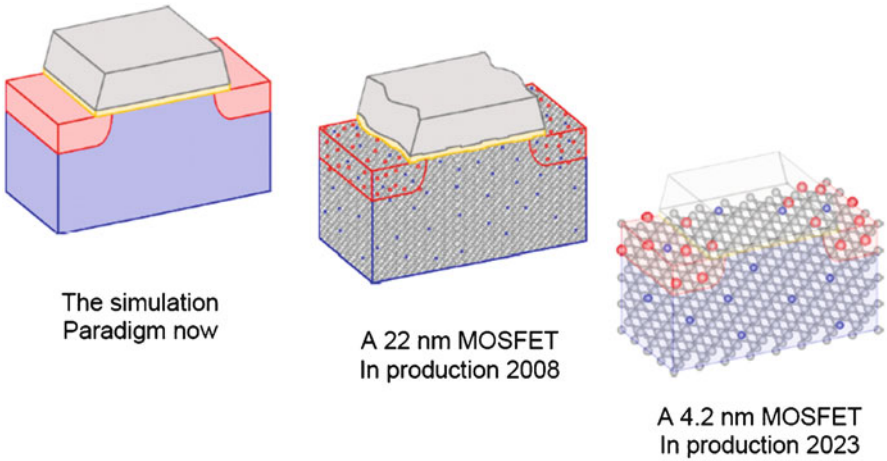


Fig. 2.1 Expected technology scaling [3]

Table 2.1 Matrix of variations

	Process	Environment	Temporal
Global	$\langle L_g \rangle$ and $\langle W \rangle$, $\langle \text{layer thicknesses} \rangle$, $\langle R \rangle$'s, $\langle \text{doping} \rangle$, $\langle t_{ox} \rangle$, $\langle V_{body} \rangle$	Operating temperature range, V_{DD} range	$\langle \text{NBTI} \rangle$ and hot electron shifts
Local	Line Edge Roughness (LER), discrete doping, discrete oxide thickness, R and V_{body} distributions	Self-heating, hot spots, IR drops	Distribution of NBTI, voltage noise, SOI V_{body} history effects, oxide breakdown currents

temporal variations. Process variations are variations due to lack of control on the fabrication process, since no two devices are exactly the same at atomic level. Environment variations are variations due to lack of perfect control over the environment (temperature, voltage, etc) in which the circuit must operate. Finally, temporal variations are variations, which cause the device to behave differently at different times, as NBTI, for example.

The types of variations can be divided in global and local. Global are variations in the value of a parameter for the entire wafer, while local are device-to-device variations within any single chip. This distinction is important because these types require different statistical treatment for proper determination of impact on yield.

With these two classifications, we can build a matrix of variations, as shown in [17]. This matrix is presented in Table 2.1.

The main focus here is about process variations, both global and local effects. In the next section we will discuss specific issues about process variations.

Figure 2.2 shows the classification of variations employed in [39]. The variations are divided into two main categories: systematic and random variations. This classification is motivated by the differences in the root causes of the different types of variation.

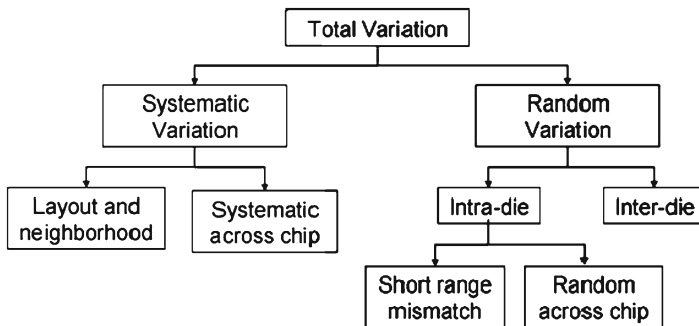


Fig. 2.2 Classification of the different types of variations in transistor characteristics [39]

According to [39], systematic variation is the difference in the electrical characteristics of two transistors with identical width and length (W/L), where there is a clearly identifiable difference in either the device layout or the layout neighborhood. Examples include the impact of gate poly pitch on gate length due to optical proximity effects, stress effects, orientation effects, etc. Random variation is the difference in the electrical characteristics of devices with identical geometry (W/L), layout, and neighborhood within the interaction distance of the known sources of variation. For example, the impact of gate polysilicon pitch differences becomes negligible for pitches greater than $3\lambda/NA$ [52], where λ is the wavelength, and NA is the numerical aperture of the exposure system. For $\lambda=193$ nm and $NA=0.75$, these give an interaction distance of approximately 775 nm. This forms a lower bound on the distance over which the local neighborhood is required to be identical for the variation to be considered random. Recently, a new source of variation with millimeter-scale interaction distance has been observed: rapid thermal anneal (RTA) temperature variation due to the density of STI regions not covered by gate poly [43, 2]. Such extreme long interaction-distance effects are considered to be random variations by [39].

Each of the main categories can be further divided. Classification of random variation into intradie (within-die) and interdie variations (between different dies) helps in identifying the root causes and possible improvement actions. For example, intradie variation can be addressed by lithography improvements, such as off-axis illumination [27] and within-field-exposure dose compensation [48]. Across-wafer nonuniformity's contribution to interdie variation can be improved by techniques like exposure dose compensation across the wafer [48], ion-implantation uniformity, RTA uniformity, etc. Systematic variation can be divided into variation arising from layout or neighborhood differences within an interaction distance.

Different aspects of circuit fabrication cause the process parameter variations. The two major sources of process variations are the fabrications steps of lithography and doping. These two will be discussed in detail.

Table 2.2 Shot noise for different energy quanta

Lithography	Energy (eV)	Resist dose (mJ/cm ²)	# quanta per 50 nm pixel	3 σ dose variation (%)
193 nm	6.4	20	500,000	0.4
EUV – 13.5 nm	92	2	3,400	5
X-Ray – 1.3 nm	920	40	6,800	4
E-beam	50,000	150 (3 μ C/cm ²)	470	14
Ion-beam	100,000	50 (0.5 μ C/cm ²)	78	34

2.1.1 Lithography

The lithographic sources of variations are the cause of both global and local variations. Imperfect lithographic process control and errors in alignment, rotation and magnification are the problems that lead to global variations. The critical dimensions are sensitive to focus, dose (intensity and time), resist sensitivity (chemical variations) and layer thicknesses [4, 9].

The local lithographic variations are due to pattern sensitivity (interference effects from neighboring shapes), interference effects from buried features, and LER (Line Edge Roughness) [10].

LER, which is primarily a discreteness effect, is due to sources of statistical variation in chemically amplified resists. These variations include fluctuations in the total dose due to finite number of quanta (shot noise), fluctuations in the photon absorption positions, nanoscale non-uniformities in the resist composition, statistical variations in the extend of acid-catalyzed deprotection, and statistical effects in polymer chain dissolution. For example, Table 2.2 shows the estimated dose uncertainty for a 50 nm contact hole, with different lithographic processes.

2.1.2 Doping

The doping process causes mainly local variations, as there are less and less dopant atoms in the transistor channel, in every new technology node. Working with few atoms can lead to a strong variability in the threshold voltage, which is proportional to the square root of the number of dopant atoms.

As showed in [15], the number of dopant atoms in the depletion layer of a MOSFET has been scaling roughly as $L_{\text{eff}}^{1.5}$. Statistical variation in the number of dopants, N , varies as $N^{-1/2}$, increasing V_T uncertainty for small N . And also, specific V_T uncertainties depend on the details of the doping profiles. Figure 2.3 shows a graphic of the scaling of number of dopant atoms versus scaling.

This doping uncertainty has a huge impact in the threshold voltage variability. Frank [16] shows an experiment with V_T measurements on 3481 identical SOI nFETs, all of single experimental macro on a single wafer (Fig. 2.4). The standard deviation can be up to 10 % of the mean threshold value.

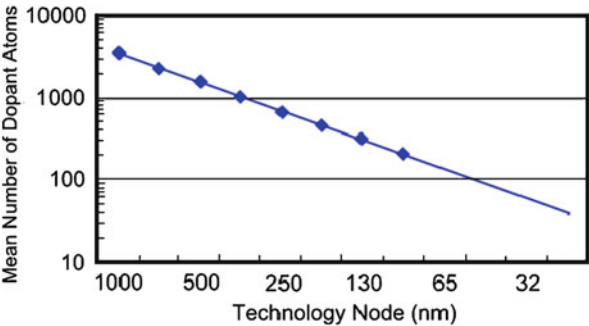


Fig. 2.3 Number of dopant atoms versus scaling [8]

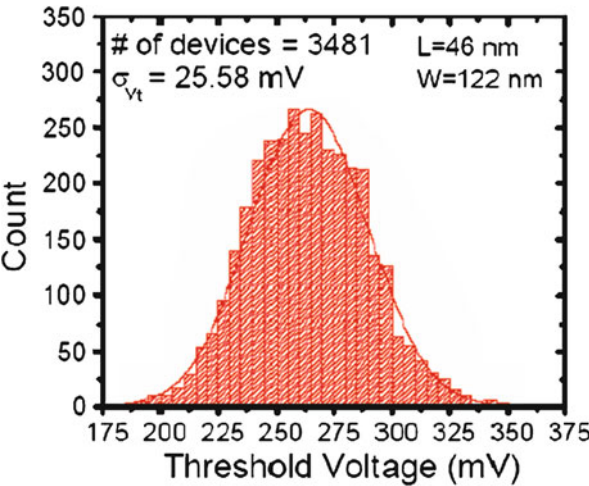


Fig. 2.4 Threshold variability [15]

2.2 Impact on Circuit Design

Process variations and intrinsic device variations cause logic and memory yield loss. Since the variations are Gaussian, and not bounded, it is not possible to absolutely guarantee functionality. Therefore, circuit designers must learn how to cope with variability, and the design must be based on achieving a target yield (90 %, 98 %, etc). The design automation tools must have means to evaluate correctly the yield.

Today, EDA tools evaluate the designs using corners. Designers usually simulate using nominal parameter values, worst-case values, and best case values. And they attempt to achieve high yield at the worst case. This is clearly too pessimistic, as

Fig. 2.5 Impact of number of critical paths on the frequency distribution [8]

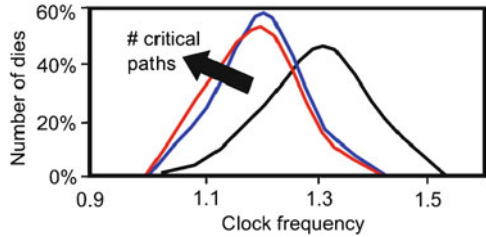
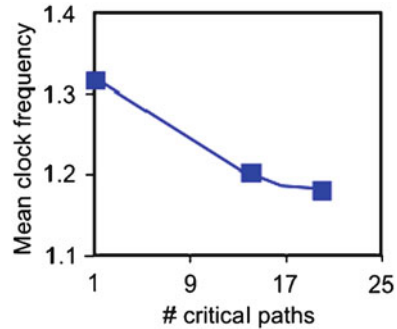


Fig. 2.6 Impact of number of critical paths on the mean frequency [8]



with process variations, it is very rare that all the devices will simultaneously show this worst-case value. It is much more possible that some of them go to one direction and others to another. This leads to statistical timing analysis tools. Now, the paths do not have one deterministic timing result, but one statistical result.

According to [37], it is necessary to verify the product response to all variations. The main responses are in power and frequency. Power variability includes variation in IDDQ (Vth fluctuation, temperature spread, voltage drop) and AC power spread (device capacitance, metal capacitance), while frequency variability includes Lg variations, thermal variations, voltage variations, and metal thickness variations. Another important product response is the possibility of failure due to setup time and hold time violations, which are a consequence of delay variability and clock skew.

In general, the fabricated circuits will show a Gaussian range of performance. Some will be so slow that will not be sold, while others will be very fast, but also too leaky. The ones near the mean value will be the good ones that go to the market.

Another serious issue related to variability and circuit design is the performance loss with the increase of the number of critical paths. As they are uncorrelated, with many critical paths, the probability that only one becomes slower than the nominal case is larger. Figures 2.5 and 2.6 shows the impact of the number of critical paths on performance.

These are the main problems that can be seen in circuit design due to process variability. It is important to develop techniques to cope with these problems, to achieve better performance and higher yield, increasing profit.

2.3 Design Techniques with Process Variations

In recent technologies, the circuits are very susceptible to yield loss. In order to cope with this loss, the design must use techniques to address yield. Currently, with the development of technology, the design challenge is changing from defect-limited yield to parametric-limited yield, which is the effect of process variations.

Since the variations are Gaussian, and not bounded, it is not possible to absolutely guarantee functionality. Therefore design must be based on achieving a target yield (e. g., 90 %). But how to evaluate the expected yield is a design automation challenge. However, if this is not taken into account in the design, there will be a loss of performance, which increases in each technology generation, as shown in Fig. 2.7. So, sooner or later, it will be necessary that all designs must be variability-aware.

This section will present three different possible design techniques to cope with process variations: Monte Carlo simulations, corner analysis, and statistical static timing analysis.

2.3.1 Monte Carlo Simulations

Monte Carlo simulation is the simplest method to predict yield, but it is also the most time consuming. It is based on a series of hundreds of normal simulations, with each simulation randomly using different electrical parameters for the transistors and the other elements that are subject to variations. These parameters can vary in a Gaussian form, or also in any other appropriate form.

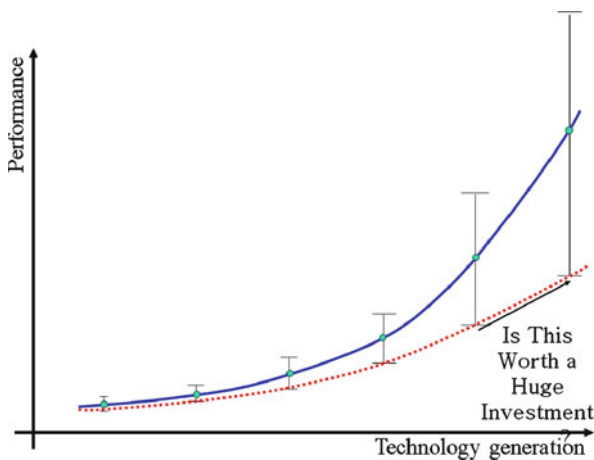


Fig. 2.7 The march of technology

After the Monte Carlo simulation is complete, it is necessary to analyze the results, and calculate the percentage of simulations that achieved the desired results, and that is the yield. However, if a high enough yield is not achieved, it is necessary to change the design, and repeat all the simulations again.

Since the simulations are very time consuming, this method can not be applied to very large designs, only to designs with a few dozen gates. However, it can also be useful to use it to single gates, and use the results in the following methods: corner analysis and statistical static timing analysis.

2.3.2 *Corner Analysis*

Corner analysis is a traditional approach that ensures good yield, at the expense of a pessimistic design, and also it is not possible to know the distribution of the yield or performance, only that the yield will be very high. A detailed explanation of corner analysis can be found in [31].

In the corner analysis, all electrical parameters are considered to be within certain bounds min and max. When they are Gaussian distributions, these bounds are normally considered to be ± 3 -sigma. If, during chip design, one finds that the circuit meets the performance constraints for all corners (boundaries), then the design is considered acceptable. Corner analysis has been applied for many years to design chips that are robust to variability.

In the corner-based analysis method, the circuits are design so that the functionality is guaranteed at all worst-case corners. The values of parameters are considered deterministic, and no probability is taken into consideration. However, the drawbacks of this method are that it makes sense only if sources of variations are strongly correlated (worst case occurs for all parameters), and it is clearly too pessimistic for deep sub-micron technologies, where there are more random variations than systematic variations.

2.3.3 *Statistical Static Timing Analysis (SSTA)*

With the increasing variability in the manufacturing process, corner analysis has been considered as inadequate and overhaul, and statistical analysis has been proposed as an alternative approach. Specifically statistical static timing analysis (SSTA) has been proposed as an alternative to traditional static timing analysis (STA) [31].

Most SSTA methods proposed use propagation of distributions of delay through the logic network. Considering the signal arrival time as a Gaussian random variable, the method consists in the propagation of the pdf (probability density function) of the random variables through the logic network, as illustrated by Fig. 2.8. Path-based SSTA does this for a given path, while block-based SSTA generates the pdf for the maximum delay of the block [31].

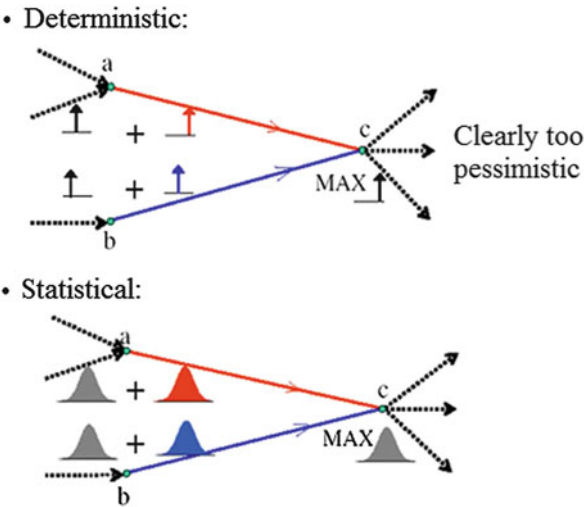


Fig. 2.8 Block-based Statistical Timing Analysis

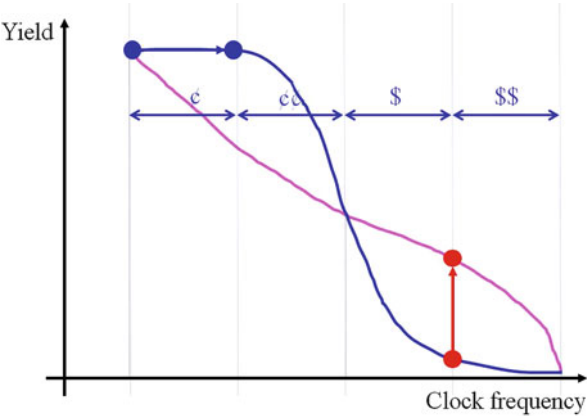


Fig. 2.9 Parametric yield curves showing profitability

The current available SSTA tools propagate the Gaussian random variables, and the final results can be used to determine the yield. Knowing the final distribution can be very useful to analyze different scenarios and determine which one is more appropriate to maximize yield (or profitability). For instance, consider Fig. 2.9. One curve is corresponding to a design with a narrow distribution of final performance. It has a high yield, since not many circuits are so slow that cannot be sold for a profit. However, it also does not have dies that will be very fast, and that can be sold for a much higher profit. When considering a wider distribution, a more significant number of dies will have to be discarded, but also many more dies will be very fast and sold by a very higher price, and the final result will be a higher profit

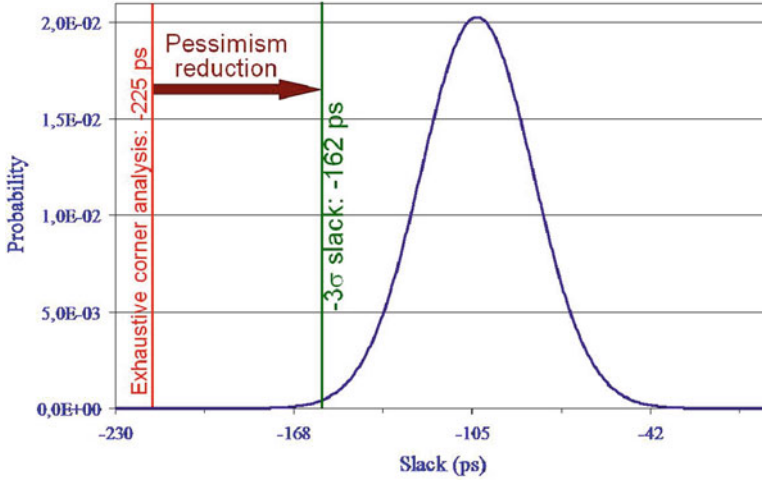


Fig. 2.10 Corner-based versus statistical timing on ASIC part

for the company. So, it is not easy to know which distribution is more appropriate for each case.

There are different tools showing the advantages of SSTA over corner analysis. For example, [49] obtains almost a 30 % improvement in the slack prediction over corner analysis, with the added advantage of knowing the exact final distribution, as presented in Fig. 2.10.

Protecting Chips Against Hold Time Violations Due to
Variability

Neuberger, G.; Wirth, G.; Reis, R.

2014, XI, 107 p. 76 illus., 51 illus. in color., Hardcover

ISBN: 978-94-007-2426-6