



Investigation of 65 nm CMOS transistor local variation using a FET array

Y.Z. Xu *, C.S. Chen, J.T. Watt

Altera Corporation, 101 Innovation Dr, San Jose, CA 95134, USA

ARTICLE INFO

Article history:

Received 8 December 2007

Received in revised form 23 May 2008

Accepted 4 June 2008

Available online 14 July 2008

The review of this paper was arranged by Prof. S. Cristoloveanu

Keywords:

Process variation

Mismatch

SRAM

ABSTRACT

CMOS FET local variation has been investigated using a new FET array structure. Key findings include four aspects. (1) At deep sub-micron technology node, local variation is significantly higher than global variation. Only 5–10% of total variation is a result of global variation. (2) Sample size affects point estimate of local variation. Sample size error can account for a significant portion of the fluctuation in the point estimate of local variation. (3) Well proximity effect (WPE) has a small impact on V_t local variation. Its impact on local variation of drive current is more significant. (4) Local variation reduces with temperature. The magnitude of NMOS V_t local variation reduction is more pronounced than PMOS. These results form a solid foundation to accurately model MOSFET local variation.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Process local variation is defined as the parametric changes of identical MOSFETs across a short distance, while global variation refers to such changes for identical MOSFETs separated by a longer distance or fabricated at different time. Normally, the local variation is within a die and the global variation is from die to die, wafer to wafer and lot to lot. The impact of local variation on circuit performance becomes increasingly important for technology nodes below 90 nm [1,2]. The primary reasons behind this trend are the continued scaling of transistor gate dimensions and the reduced pace of gate oxide thickness scaling [3,9]. Additionally, new processes have been introduced to contain short channel effect which may negatively impact local variation. For example, more nitrogen is incorporated into the gate dielectric to reduce boron penetration and low thermal budget is used to improve short channel effects. These new processes may add more variation to FET characteristics. It is widely believed that the local variation will directly impact yield for circuits such as SRAM for technology nodes below 65 nm. New circuit techniques will be needed to work around the strong impact from local variation on circuit functionality and parametric yield. Accurate characterization and modeling of local variation is imperative for simulations to accurately predict the yield and evaluate the benefit of the new circuit techniques. To achieve accurate results, the local variation needs to be accurately modeled at the process, temperature and voltage conditions for which circuit yield is most affected.

Traditionally, MOSFET mismatch is measured by placing two transistors side by side. The difference of MOSFET parameters such as V_t , I_{dsat} and G_m between two transistors is recorded and the standard deviation of the difference measured over many pairs is used as an estimate of the local variation. The standard deviations follow a linear dependence with respect to the reciprocal of the square root of the channel area [4]. Due to silicon area limitations, the number of transistor pairs inside one reticle frame is limited. Statistical significance is achieved by pooling data from multiple dies and wafers. As a result, the data includes contributions from both local variation within a die and fluctuations of the local variation from die to die. The measured local variation values can change from wafer to wafer and lot to lot. In recent years, test structures employing a FET array to characterize local variation have been proposed [5]. However, periphery circuits are needed to access individual transistors and at early stages of process development, such circuits may not be readily available.

In this paper, we report the results of a FET array structure for accurate characterization and modeling of local variation. The FET array is laid out in a way such that individual transistors can be accessed without periphery circuits. Metal layers are optimized, so that minimum IR drop (the product of the current (I) and the resistance (R) of a conductor) occurs in the test structure. Using the structure, the local variation of transistor parameters is accurately measured. The measurement is also carried out at different temperatures. Well proximity effect (WPE) impact on local variation is measured using the test structure as well.

* Corresponding author. Tel.: +1 408 544 7387; fax: +1 408 544 7594.
E-mail address: yxu@altera.com (Y.Z. Xu).

2. Test structure and measurement results

A schematic of the FET array is shown in Fig. 1. Transistors in the same column share a common drain bus, while each row of transistors share a common gate bus. All transistor source and bulk nodes are tied together. To access each individual transistor, proper biasing of the drain and gate voltage is needed. An example of measuring FET T11 is shown in the figure, in which column one is biased while the rest of the drain columns are left open. The voltage of the first gate row is swept to find the threshold voltage of T11 while the rest of the gate rows are kept at 0 V or slightly less than 0 V to minimize the leakage in column one, which impacts measurement of V_t . All drain columns have been strapped with multiple metal levels so that the corresponding IR drop is negligible. It is noteworthy that the test structure is not suitable for leakage measurement.

The test structure was manufactured using a 65 nm low power process [7]. Each array includes 100 FETs. FET W/L is $0.6 \mu\text{m}/0.065 \mu\text{m}$. Dummy active area and poly features are placed around the array to minimize proximity effects. At one edge of the array, an N-Well or P-Well feature is placed at the minimum allowed design rule, so that FETs at one side of the array receive the most severe WPE. This allows a direct comparison of WPE impact on transistor local variation.

2.1. Local and global variation

The measured V_t follows a normal distribution as shown in Fig. 2. The data plotted in the figure includes all FETs within many identical arrays across the wafer. From the same data, the median V_t for each array is plotted as a 2D contour plot in Fig. 3. It is clear that V_t is rather uniform across wafer. Local variation of different types of transistors are then extracted from the measured data and shown in Table 1.

The average, maximum and minimum standard deviation represents site to site variation across a wafer. In Table 2, the percentages of drive current standard deviation over average current are listed. NMOS local variation is higher than PMOS. Local variation of transistor characteristics is mainly due to dopant fluctuation, but other process, such as gate oxide variation, interface charge and poly-gate, line edge roughness (LER) [8], may also increase the variation.

Total, local and global variation follow the equation below

$$\sigma_{\text{total}}^2 = \sigma_{\text{local}}^2 + \sigma_{\text{global}}^2 \quad (1)$$

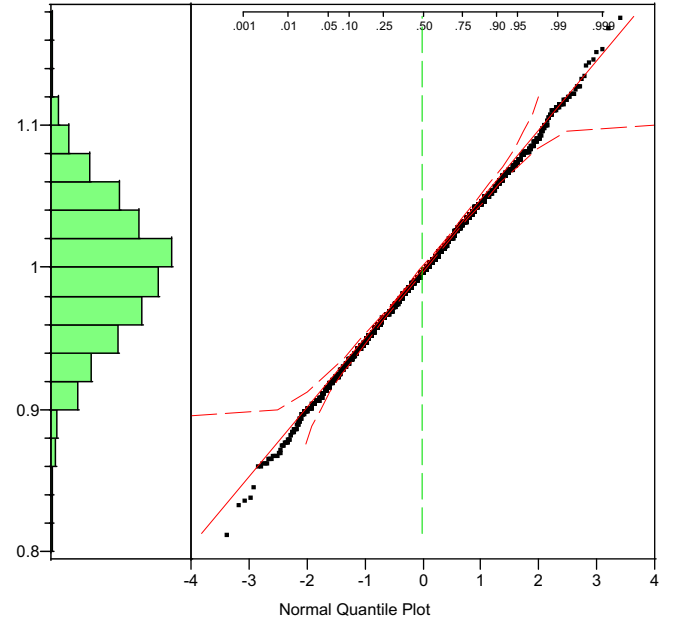


Fig. 2. Normalized V_t distribution of all n FETs across wafer.

Here σ_{total} , σ_{local} and σ_{global} represent the standard deviation of total variation, local variation and global variation respectively. The global variation includes components of die-to-die, wafer-to-wafer and lot-to-lot variation.

Using the measured data, we can separate local variation from global variation using the nested variation approach. The variance of the average value measured for each array comprises both local variation and global variation. Suppose there are n FETs in a local array and iD instances of the array, then the nested variances follow the following equations

$$\frac{\sum_{j=1}^D \sum_{i=1}^n (x_{ji} - \bar{x}_j)^2}{(n-1) \cdot D} = \hat{\sigma}_{\text{local}}^2 \quad (2)$$

$$\frac{\sum_{j=1}^D (\bar{x}_j - \bar{x})^2}{(D-1)} = \hat{\sigma}_{\text{global}}^2 + \frac{\hat{\sigma}_{\text{local}}^2}{n} \quad (3)$$

Here $\hat{\sigma}$ is the point estimate of standard deviation [10]. The calculated partition of local and global variation is shown in Table 3. The results indicate that within one wafer, the local variation dominates for this process.

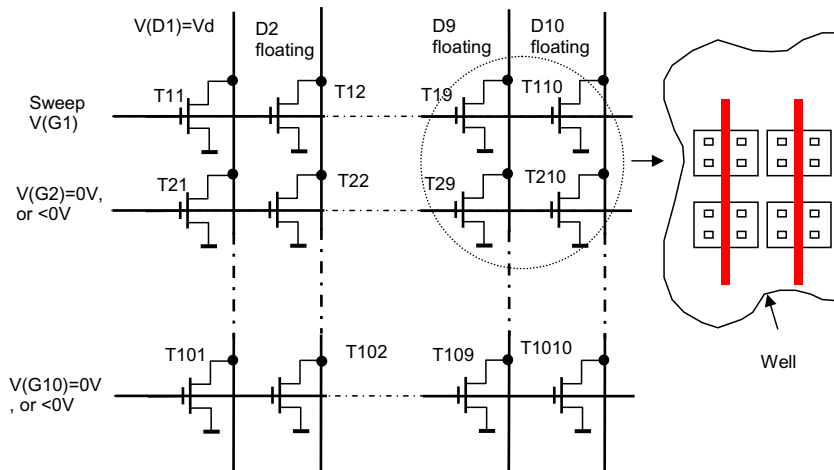


Fig. 1. Schematic of the FET array. Inset shows a corner of layout featuring one sided well proximity effect.

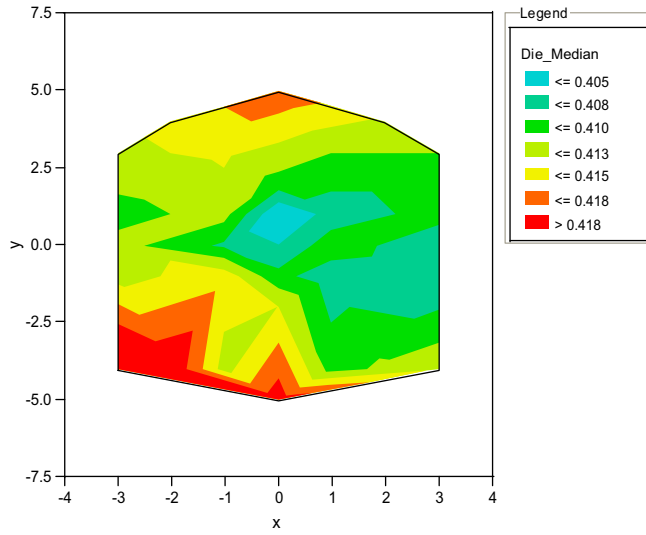


Fig. 3. Distribution of NMOS median V_t of each die across a wafer. X and Y axis are frame index number of each die. Wafer center is (0,0).

Table 1
 V_t local variation of different types of FETs

FET type	Average σV_t (mV)	Max σV_t (mV)	Min σV_t (mV)
NMOS	19.7	22.2	16.9
PMOS	12.8	19.1	10.8

Table 2
 I_{dsat} local variation for NMOS and PMOS

FET type	Average σI_{dsat} (%)	Max σI_{dsat} (%)	Min σI_{dsat} (%)
NMOS	3.3	3.9	2.9
PMOS	2.9	3.3	2.4

Table 3
Partition of V_t local and global variance and ratio

FET type	Local variance (mV) ²	Global variance (mV) ²	Ratio of global/local
NMOS	388.5	15.4	4%
PMOS	162.8	9.0	6%

2.2. Impact of sample size on local variation results

As can be seen from Tables 1 and 2, the local variation has considerable variation from die to die. It is useful to investigate if the variation is due to process variation or if it arises from sampling induced error. With the given sample size, we calculate the possible variation of standard deviation at a set confidence level.

When sample size is large, the point estimate of standard deviation approaches the true standard deviation. However, when sample size is small, a significant difference between the true standard deviation and point estimate exists. For a point estimate of σ based on n random samples of a normal distribution, variance s^2 is associated with a χ^2 distribution [6] following Eqs. (4) and (5).

$$\chi^2 = \frac{s^2 \cdot (n-1)}{\sigma^2} \quad (4)$$

$$P(\chi^2) = \text{constant} \cdot \left[(\chi^2)^{\frac{(n-1)}{2}-1} \right] \cdot e^{-\frac{\chi^2}{2}} \quad (5)$$

The χ^2 distribution, which is a skewed one, has $n-1$ degrees of freedom. $P(\chi^2)$ is the corresponding probability density. S^2 repre-

sents the sample variance with a limited sample size n . The point estimates of variance are compared with the theoretical χ^2 distribution in Fig. 4. The measured data follows the theoretical cumulative distribution.

The result here shows the point estimate of standard deviation can vary for a given sample size n . With increase of sample size, the slope in Fig. 4 becomes steeper. Therefore, there is less variation for the point estimate. In theory, the sample size should have a similar impact on the global variation. In this case, the sample size is the number of die from which one collects total variation information, not the number of FETs in the array. As mentioned previously, the analysis of global and local variations is based on a nested variation method. The accuracy of local variation affects the derived results of the global variation. Additionally, the variance χ^2 distribution is based on the assumption that V_t and I_{dsat} follow the normal distribution. We have validated such assumption in Fig. 2. When employing the described method for the sample size impact on global variation, a similar verification should be done.

2.3. WPE and temperature impact on local variation

Well proximity effect results in increased transistor V_t and reduced drive current for transistors close to a well boundary. This is due to increased channel doping caused by the scattering of implanted ions from the photoresist edge during well formation. Since random dopant fluctuations are one of the major causes of transistor mismatch or local variation, it is possible that WPE may impact local variation. As described previously, the FET array in our test structures can be used to assess WPE impact on local variation. The transistors on one edge of the array are close to a well boundary; all other transistors are far from a well boundary. The minimum design rule is used to place the well boundary adjacent to FET array on one side. This creates the worst-case one-sided WPE for the transistors on this side. Table 4 shows the measured V_t and I_{dsat} local variation for the transistors adjacent to the well boundary.

Compared to the local variation of FETs without WPE effect, both NMOS and PMOS local variation reduces by 4% on average. Due to the smaller sample size of FETs close to the well boundary, the variation of the standard deviation is larger than FETs without WPE. This agrees with the theoretical calculation made in Section 2.1. The impact of WPE on drive current local variation can be derived by comparison to the previous table as well. The average drive current local variation increases by 39% and 10% for NMOS and PMOS respectively. These results suggest that drive current local variation degrades more significantly than V_t local variation due to WPE. Using the method described previously, the local variation and global variation can be partitioned for the FETs impacted by WPE. The ratio between die-to-die variance and local variance are comparable to the results of FETs without WPE.

The effect of temperature on local variation is important for accurate circuit simulation. Local variation has greatly reduced the design margin for some circuits such as SRAM. The local variation of the transistors within the bitcell can significantly degrade the read margin and write margin for the worst-case cell within a large array. Since the design margin is typically a function of temperature, it is important to correctly account for any temperature dependence of local variation in analysis of circuit functionality and yield. For some analysis such as SRAM read margin the worst-case condition may be high temperature, while for other analysis such as SRAM write margin the worst-case condition may be low temperature. It is therefore important to characterize the local variation across temperature. In this study, the transistor arrays on some die were measured both at room temperature and high temperature. The corresponding local variations are plotted in the same figure. Figs. 5 and 6 show the temperature impact on V_t

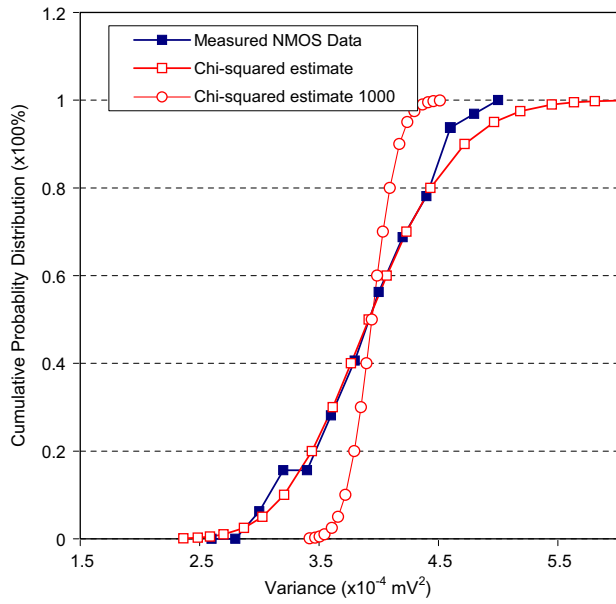


Fig. 4. Comparison of cumulative distribution of a χ^2 distribution and measured variance. Curve with label of 'Chi-squared estimate 1000' shows the cumulative distribution of variance of sample size 1000.

Table 4
Local variation of FET receiving one-sided WPE

FET type	Avg σV_t (mV)	Max σV_t (mV)	Min σV_t (mV)
<i>V_t local variation</i>			
NMOS	18.9	25.0	11.7
PMOS	12.3	17.4	7.5
FET type	Avg σI_{dsat} (%)	Max σI_{dsat} (%)	Min σI_{dsat} (%)
<i>I_{dsat} local variation</i>			
NMOS	4.6	6.7	2.8
PMOS	3.2	4.8	1.7

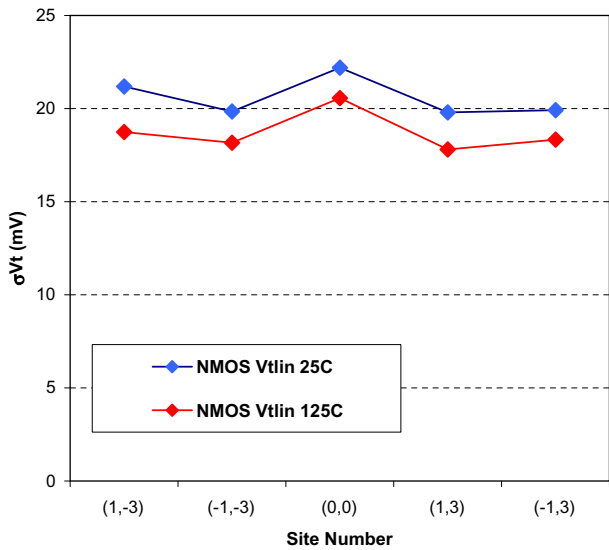


Fig. 5. NMOS V_t local variation across five sites at 25 °C and 125 °C.

local variation, while the impact on NMOS and PMOS I_{dsat} are shown in Figs. 7 and 8 respectively.

It is clear that local variation is reduced at high temperature. The average magnitude of local V_t variation reductions are 10%

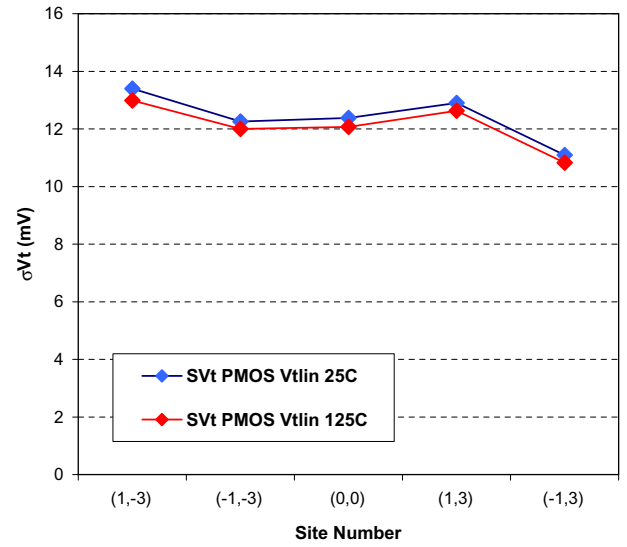


Fig. 6. PMOS V_t local variation across five sites at 25 °C and 125 °C.

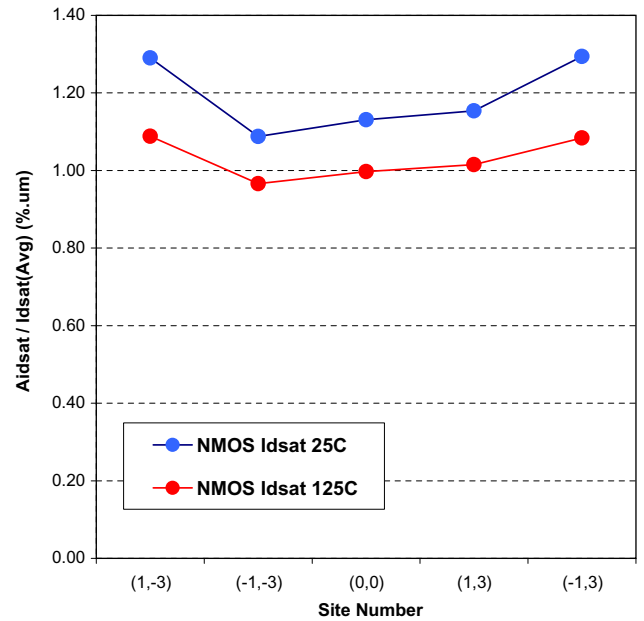


Fig. 7. NMOS I_{dsat} local variation across five sites at 25 °C and 125 °C.

and 3% for NMOS and PMOS respectively. I_{dsat} local variation has the same trend, but the magnitude of the local variation reduction is more than 10%. It is believed that the reduction in V_t local variation at high temperature is caused by the reduction in channel depletion region depth. This reduces the number of ionized dopant atoms in the channel and hence reduces their fluctuation. The difference between NMOS and PMOS local variations may arise from the contributions of other components of local variation such as LER, gate dielectric charge or dielectric thickness variation. The change in local variation at high temperature is significant and should be accounted for in any analysis which is sensitive to local variation. Although low temperature local variation behavior was not measured in this study, it is expected that local variation will increase if temperature is reduced below room temperature due to the widening of the channel depletion region.

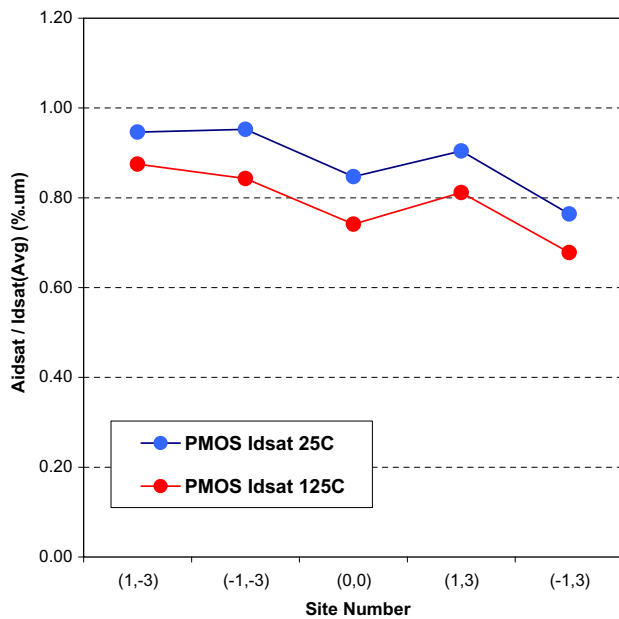


Fig. 8. PMOS I_{dsat} local variation at 25 °C and 125 °C.

3. Conclusion

CMOS transistor local variation has been fully characterized by using a FET array structure. The higher statistical significance generated by a relatively large number of FETs in a local area provides a foundation to accurately measure and separate local and global variation. Measurement results indicate that at deep sub-micron technology nodes local variation is significantly higher than global variation. Our results show only 5–10% of the total variation arises from die-to-die variation, while the majority comes from local variation. Using statistical theory, the impact of sample size to the variation of the point estimate of standard deviation derived from the FET array is studied. It is found that the fluctuation in the measured

local variation caused by the limited sample size of transistors within the array accounts for a significant portion of variation of the point estimate of standard deviation from die-to-die. As a result, care must be taken to correctly interpret the measured results and avoid using an overly pessimistic estimate of local variation. WPE impact on local variation has also been characterized. It is found that WPE causes a small change of V_t local variation. The change of drive current local variation due to WPE, however, is more significant than V_t local variation. Finally, the temperature dependence of local variation has been characterized by comparing the measurement of local variation at room temperature and high temperature. Test results reveal an unequivocal signal of variation reduction with temperature. The magnitude of NMOS V_t reduction is more pronounced than PMOS.

Acknowledgements

The authors would like to thank Clement Hsu, Chris Pass, Dale Ibbotson for help and useful discussions. We would also like to thank Chung-Kai Lin, Sally Liu, C.K. Yang, Shien-Yang Wu, H.C. Tseng and Michael Wang from TSMC for their help and discussion on local variation on temperature dependency.

References

- [1] Pille J, Adams C, Christensen T, Cottier S, Ehrenreich S, Kono F, et al. ISSCC 2007:322.
- [2] Wang Y, Ahn H, Bhattacharya U, Coan T, Hamzaoglu F, Hafez W, et al. ISSCC 2007:324.
- [3] International technology roadmap for semiconductors. 2006.
- [4] Pelgrom MJM, Duinmaijer ACJ, Welbers APG. IEEE J Solid State Circ 1989;24(5):1433. October.
- [5] Quarantelli M, Saxena S, Dragone N, Bacock JA, Hess C, Minehane S, et al. ICMTS 2003:238.
- [6] Devore Jay L. Probability and statistics for engineering and the sciences. Brooks/Cole Publishing Company; 1982.
- [7] Fung S, Huang HT, Cheng SM, et al. VLSI Symposium 2004:92.
- [8] Diaz CH, Tao HJ, Ku YC, Yen A, Young K. IEEE Electr Dev Lett 2001;22(6):287.
- [9] Roy G, Brown AR, Adamu-Lema F, Roy S, Asenov A. IEEE Tran-ED 2006;53(12):3063.
- [10] P Box GE, Hunter WG, Hunter JS. Statistics for experiments. Wiley Interscience; 1978.