Using TCAD, Response Surface Model and Monte Carlo Methods to Model Processes and Reduce Device Variation

Dipanjan Basu
Microelectronics Research Center,
The University of Texas at Austin,
10100 Burnet Road, Bldg. 160, Austin, TX 78758, U.S.A.
dipanjan@mail.utexas.edu

J. Guha, P. Hatab, P. Vaidyanathan, and C. Mouli.
Micron Technology Inc., Boise, ID, USA.

S. K. Groothuis.
Consultant, SimuTech Group.

Abstract— **Reduction of electrical parameter variation is essential to achieve high yield and reliability in semiconductor devices. However, variation depends on a large number of process factors, which are often interdependent. In this work, well-calibrated Technology Computer-Aided-Design process and device simulations were performed in a designed experiment to develop an efficient, surrogate response surface model (RSM) of the device parameters as a function of key process factors. Monte Carlo simulations were performed with the RSM to estimate variation and design systems to reduce variation. The approach, illustrated here specifically for peripheral n-type field-effect transistors in a dynamic random-access-memory process flow, is general, easy-to-implement, and a cost-effective way to systematically identify, model, and analyze process variation.**

Keywords-component; response surface; process variation

I. Introduction

Scaling has resulted in increasingly complex and difficult-to-control unit processes, and the ensuing variation in electrical parameters has become a key issue in maintaining device yield and reliability. Variation due to the discrete nature of matter is fundamental; however, extrinsic variation, due to deviations in processing conditions of multiple factors may be reduced [1]. A systematic analysis of process variation requires a large number of experiments, and it becomes prohibitively expensive, even with an optimized design, if one wishes to use real Si data. This cost can be reduced by using process and device simulations from properly calibrated Technology Computer-Aided-Design (TCAD) software [2]. However, the computational load remains very high and analytic functions developed to serve as surrogate models are best suited for a cost-effective study. Several such integration schemes with different methods and scopes of operation have been proposed [3-6].

For illustrating the methodology in this work, we chose peripheral transistors in the stacked-capacitor dynamic random-access-memory (DRAM) process flow. Here, the processes are optimized to improve the retention time of memory cells [7]. For simplicity, in this article we restrict the target electrical parameter to threshold voltage (V_T) only, and show results from n-type MOS (nMOS) field-effect transistors (FET) having channel length (L_{ch}) of 95 nm.

II. TCAD Process and Device Simulations

A. *Input process factors for screening*

The output electrical parameter (here V_T) is, in general, a non-linear function of the input process parameters. This non-linearity can be modeled by a second order polynomial if the PF variation remains small, (within 5-10% of nominal value). The parameters mostly vary in a normal distribution, and metrology data, wherever available, indicates that typically the $\mu \pm 3\sigma$ value falls well within the nominal ± 5−10% range ($\mu-$ mean, $\sigma-$ standard deviation). Therefore, the smaller of these two available ranges of input parameters was selected, so that within the tight realistic process bounds, the input-output model would be comparatively more accurate.

The input PFs, listed in Table I, were chosen from three broad categories – structural, thermal and doping-related. Multiple sources of variation were investigated within each of these processes, e.g. ramp-rates and final temperatures for the anneal steps, variations in dose and energy for the implantation steps. Physical insights went into choosing the parameter sets, and the control levels for the simulation, e.g., the thermal parameters comprised of two high temperature (>1000 ˚C) short time processes, anneal after borophosphosilicate glass (BPSG) deposition, and a rapid thermal process (RTP) for tetraethylorthosilicate (TEOS) film densification and activation of polySilicon plug contacts. While the final peak temperature in the single wafer spike anneal furnaces are difficult to control and measure, the nitridation processes, typical in a stacked-DRAM process to build buried digit lines and cell capacitors, are relatively well controlled, long time span (~ 1 hour), lower temperature (~ 750 ˚C) processes. However, there often exists a temperature gradient across the nitridation furnaces, resulting in an across-the-wafer temperature variation, which was accounted for as a source of variation.

B. *Screening of input PFs for multivariate study*

We employed a simple three point experiment to screen the input PFs, where the PF of interest was varied to the low (-3σ), nominal, and high (+3σ) operating point. The slope of output parameter vs. input PF was calculated as an average of the forward (high to nominal) and backward (nominal to low) slopes, to account for the non-linearity of V_T with respect to the input PF variation. For a process factor with nominal parameter value P, a normalized sensitivity parameter was defined as –

$$S = \left(dV_T/V_T\right)/\left(dP/P\right). \qquad (1)$$

The dimension-less S allowed us to uniformly compare the effect of different types of parameter such as L_{ch} and T_{BPSG}. In Fig. 1 we show the absolute value of S for two each of the different categories of PF variation. By comparing $|S|$ for the individual PF variations, we narrowed our PF space to a six-

978-1-4244-3947-8/09/$25.00 ©2009 IEEE

TABLE I. PROCESS FACTORS CONSIDERED

Parameter Category	Process Factor (PF)
Structural	Gate oxide thickness (T_{ox}) Channel length (L_{ch}) Spacer thickness (T_{spcr}) Source/Drain epi-Si thickness (T_{epi})
Thermal	Borophosphosilicate glass (BPSG) deposition Tetraethylorthosilicate (TEOS) film densification Buried digit line nitridation Cell capacitor nitridation
Implant	Halo Lightly doped drain (LDD) Source/Drain

dimensional space consisting of T_{ox}, L_{ch}, peak BPSG temperature (T_{BPSG}), halo dose (*Halo*), LDD dose (LDD_{Ds}), and LDD energy (LDD_{En}).

In choosing these PFs for a multivariate study, it is better to include PFs from all three categories, as then it becomes possible to design for systems that reduce variation by utilizing the correlation of these variables (see Section III). If one restricts to choosing the variables which give maximum $|S|$, one may end up with too many identical type of PFs in the model, e.g. for PMOS, the thermal parameters all result in high $|S|$ because of high diffusivity of B. However, this does not result in a feature-rich model useful for design.

C. Response surface model and design of experiment

Response surface models are polynomial functions that approximate a true, but complex and possibly unknowable *Response*. The degree of the polynomial required for a good fit of the data dictates the design of experiment (DoE) required for data collection. To fit a 2nd order polynomial, a DoE with at least 3 levels is required. For optimal performance, we chose a central composite design, involving $2^k + 2k + 1$ runs (k– number of PFs), as opposed to a full factorial design (3^k) [8].

Since V_T response for some PFs (e.g. T_{ox}) is linear in the narrow range of variation of the PFs, a full second order polynomial RSM is not necessary. We constructed the model in a stepwise regression manner where each of the $(k+1)(k+2)/2$ terms entered into the model if they had a Probability-to-Enter of less than 0.05, and they remained in the final model with a Probability-to-Remove of less than 0.20. This keeps the model simple by trimming non-significant factors [91].

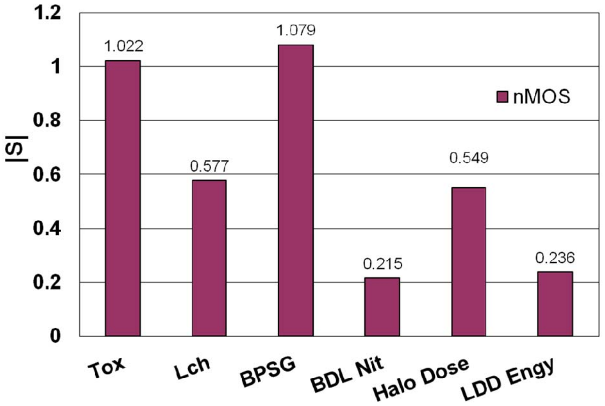

Figure1. Absolute sensitivity index for six individual process factor (PF) variation for peripheral n-type MOSFET in a standard DRAM process flow.

Table II gives the terms in our final model of V_T for nMOS devices. An excellent match of the actual V_T (TCAD) to predicted V_T (RSM) was obtained (see Fig. 2). The predicted and the actual values were usually within 5mV of each other. The goodness of the fit is estimated from the coefficient of determination (R^2). Our model gave $R^2 = 0.996$ ($R^2 =1$ for a perfect fit). As a further check, we selected several random points within our six-dimensional (6-D) PF space, and performed full TCAD process and device simulations to calculate V_T at those points, and compared the values with predictions from the model. Fig. 2 shows good match of the predicted and actual V_T for these random points (denoted by star). In practice, carefully selected random points in the PF space can serve as additional inputs to the predictor to improve the accuracy of the RSM over the entire PF space. These random points can be generated based on input PF statistic. In addition, Latin hypercube sampling can be employed instead of rudimentary truly random sampling to optimally represent the entire PF space [10].

TABLE II ESTIMATES OF COEFFICIENTS OF THE RSM TERMS

Term	Unit[1]	Value
Constant		0.0867
T_{ox}	1 nm	0.222
L_{ch}	1 nm	1.5x10^{-3}
T_{BPSG}	1 °C	-3.8x10^{-4}
Halo	10^{13} cm^{-2}	0.0862
LDD_{Ds}	10^{14} cm^{-2}	0.0268
LDD_{En}	1 keV	-0.0118
$(T_{ox} - T_{ox0})(L_{ch} - L_{ch0})$		9.12x10^{-4}
$(L_{ch} - L_{ch0})^2$		-7.51x10^{-5}
$(L_{ch} - L_{ch0})(Halo - Halo_0)$		-3.27x10^{-4}
$(L_{ch} - L_{ch0})(LDD_{Ds} - LDD_{Ds0})$		1.23x10^{-3}
$(T_{ox} - T_{ox0})(LDD_{En} - LDD_{En0})$		-9.3x10^{-3}
$(L_{ch} - L_{ch0})(LDD_{En} - LDD_{En0})$		4.08x10^{-4}
$(Halo - Halo_0)(LDD_{En} - LDD_{En0})$		-2.64x10^{-3}
$(LDD_{Ds} - LDD_{Ds0})(LDD_{En} - LDD_{En0})$		-6.56x10^{-3}
$(LDD_{En} - LDD_{En0})^2$		3.589x10^{-3}

[1] Units serve as normalization factors, e.g., for an LDD Dose of 3x10^{14} cm^{-2}, the value of LDD_{Ds} is 3. The subscript 0, as in T_{ox0} denotes the nominal or quiescent point.

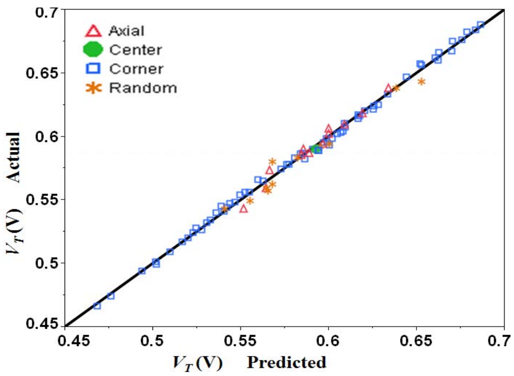

Figure 2. Actual V_T (from TCAD) vs. predicted V_T (from RSM) showing good fit of the response surface model. The center line is the perfect fit (45˚) line. The spatial location of the points in the 6-D PF space is also shown. R^2=0.996 confirms an excellent fit (R^2=1 for perfect fit).

III. Applications of RSM

A. *Estimation of variation for generating corner models*

The response surface model, being an analytic function that is accurate within the range of the input 6-D PF space modeled, can be used to quickly generate a distribution of V_T by feeding in thousands of sets of PFs from a Monte Carlo (MC) simulator. For this purpose, we used the statistical tool JMP [11]. The result is shown in Table III, where we fixed the three implant parameters to their nominal values, other PFs were distributed normally around the nominal values (standard deviation– σ of each is listed in Table III). In Fig. 3 is shown a snapshot of this calculation from JMP along with the functional dependence of V_T on the input PFs. The histogram of V_T for the baseline case (based on metrology data) is shown in greater detail in Fig. 4. We would like to mention that this is tighter (σV_T =11.6 mV) than production data, since not all sources of variation have been taken into account in the RSM.

The use of JMP allowed us to easily a) specify a variety of distributions (Gaussian, Weibull, exponential, triangular, etc.) and b) specify a multivariate correlation structure between the PFs. Correlations between PFs can exist (e.g., the authors have seen a slight negative correlation between L_{ch} and T_{spcr}) and the input MC distributions will change accordingly. In such cases, the variation of V_T can be inferred from the functional dependence curves in Fig. 3, e.g., V_T increases on increasing T_{ox}, and decreases on increasing T_{BPSG}, so that a negative correlation between T_{ox} and T_{BPSG} leads to an increase in variation of V_T. For illustrative purpose only, we assumed that there is a correlation between T_{ox} and T_{BPSG}. Keeping rest of the distributions identical to the baseline, for a correlation between T_{ox} and T_{BPSG} equal to –0.5 (+0.5) σV_T increases (decreases) to 12.1(10.1) mV from the baseline value of 11.6 mV.

The effect of tightening or loosening an input PF distribution is shown in Table III for various T_{ox}, L_{ch} and T_{BPSG}. Combining the versatile MC simulator of JMP and the fast RSM, one gets a very efficient tool to estimate device level variation due to PF variation. This approach is particularly useful to construct pre-Si corner models, with the anticipated distributions of the key PFs serving as key inputs.

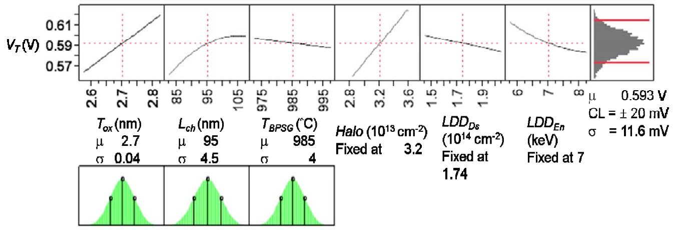

Figure 3a. Snapshot of the Monte Carlo simulation of the RSM from the JMP software, showing in top left, the functional dependence of V_T on the input PFs, and in top right, V_T distribution, when T_{ox} L_{ch} and T_{BPSG} vary according to a normal distribution (bottom left).

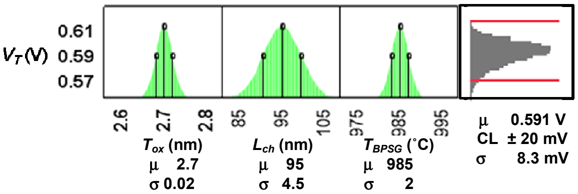

Figure 3b. Tighter distribution of T_{ox} and T_{BPSG} result in a comparatively narrow V_T distribution that is slightly asymmetrical.

TABLE III Variation of V_T due to Input PF Variation

σ for Input PFs			Output σ	Comments
T_{ox} (nm)	L_{ch} (nm)	T_{BPSG} (˚C)	V_T (mV)	
0.04	4.5	4	11.6	Baseline distribution
0.06	4.5	4	15.2	
0.06	2	1	13.5	
0.03	3	1	8.1	

B. *Optimizing PF space for low defect rate*

There are a variety of ways to search the PF space to obtain a quiescent point that matches one or more preset criteria and yet is comparatively insensitive to input PF variations [12]. In this work, our goal is to keep V_T within a range of the target V_{T0}; therefore, we defined a defect rate (DR), which estimates the percentage of V_T that lie outside the defined upper and lower specification limits (USL and LSL respectively, see Fig. 4), selected as +/– 20 mV from V_{T0}. For the baseline process, DR turned out to be 8.88%.

A designed experiment was then performed, where the entire 6D PF space was covered, using Latin hypercube sampling to have an efficient, representative coverage. At each of these 128 points, 10^4 MC simulations were run to obtain the DR. Subsequently, the overall DR surface was approximated by a Gaussian Process model, a popular fitting technique for computer simulation, where each point on the DR surface is predicted from a weighted average of the neighboring points [13]. The minima of this surface (expressed in Table IV in terms of change of PFs from the nominal values) gave us the operating condition that minimized DR. For this optimized point, DR went down to 4.42%, σV_T reduced to 9.7 mV. The construction of the DR surface through the Gaussian Process model and finding minima on that surface ensures that one gets the true local minima of the PF space, and is not restricted to the sample points of the initial LHC design experiment.

The DR approach has certain advantages over modeling and minimizing σV_T. It allows for asymmetric specification limits for situations where, for example, occurrence of a low V_T

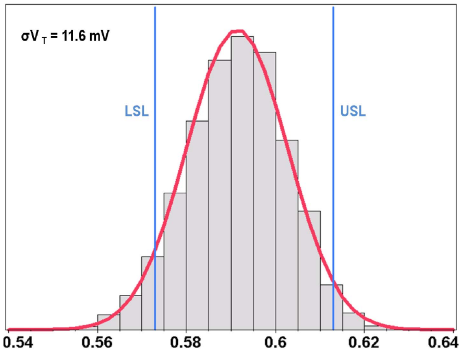

Figure 4. Baseline distribution of V_T showing lower and upper specification limits. T_{ox}, L_{ch} and T_{BPSG} follow normal distribution (see Table III).

978-1-4244-3947-8/09/$25.00 ©2009 IEEE

TABLE IV. Movement of the PF Parameters from the Nominal Values To Obtain a Low Dr Operating Point

Term	Unit[1]	Change of PF values from the nominal value (baseline)	σ
T_{ox}	1 nm	-0.03	0.04
L_{ch}	1 nm	+7.2	4.5
T_{BPSG}	1 ˚C	+2.4	4
Halo	10^{13} cm^{-2}	-0.24	
LDD_{Ds}	10^{14} cm^{-2}	+0.24	
LDD_{En}	1 keV	-0.89	
V_T	1 mV		9.6

is more troublesome than a high V_T. The DR used in this study was a step function (0 if within the spec limits, 1 if outside). A DR based on a customized Loss Function could also be employed. It is worthwhile to note that, in practice, careful consideration of the impact of the process change on other aspects of device performance and reliability should accompany any optimization.

C. Design of a feed forward system

To keep V_T at a target value (V_{T0}), a system can be developed to feed forward upstream process variation, e.g., that in T_{ox} and L_{ch}, to modify a downstream process such as B Halo dose (*Halo*). To illustrate how upstream metrology information could be used to adjust the B Halo dose to correct for the effects of an off-target T_{ox} and L_{ch}, we modeled *Halo* as a function of predicted V_T from the RSM in absence of halo dose correction, T_{ox} and L_{ch} to obtain:

$$\begin{aligned} Halo &= 4.79 - 2.28T_{ox} - 0.014L_{ch} + 10.24V_T \\ &- 0.03(T_{ox} - T_{ox0})(L_{ch} - L_{ch0}) + 0.11(L_{ch} - L_{ch0})(V_T - V_{T0}) \end{aligned} . \quad (2)$$

For the 5000 points in the baseline data set, if the halo dose was adjusted (for each pair of off-target T_{ox} and L_{ch}) according to (2), we would get an extremely narrow V_T distribution, with σV_T of only 3 mV, and DR = 0. This is demonstrated in the V_T distribution in Fig. 5.

This process demonstrates the efficacy of a simple feed-forward system to reduce variation in output parameters. However, it should be noted that this analysis assumes each V_T could be adjusted at the die-level, whereas in practice, the adjustment would probably take place at the wafer level, and therefore, would be less effective.

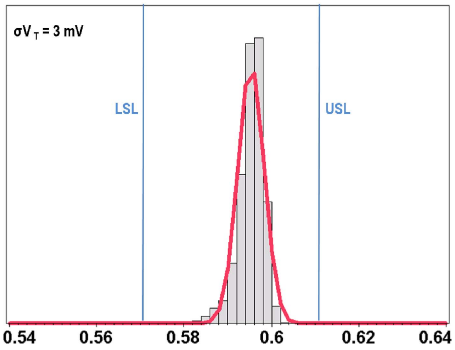

Figure 5. Distribution of V_T for the same input PFs as in the baseline case (Fig. 4), except the halo dose (Halo), which has been adjusted through feed forward design. Adjustment of halo dose for deviations of T_{ox} and L_{ch} results in a very narrow distribution.

IV. Conclusion

Properly calibrated TCAD simulations were used to build a simple, analytic response surface model to serve as a surrogate model to estimate the effect of process variation on V_T of n-type MOSFETs in a DRAM process flow. A simple and intuitive normalized sensitivity index was used to identify the PFs which cause maximum V_T variation. Monte Carlo simulations were performed on the surrogate model to estimate variations in V_T for different input PF distributions and optimize process factors to reduce variation. A simple feed-forward model developed using regression analysis, exhibits the benefit of that can be gained from design and analysis with a representative analytic model.

Acknowledgment

D. B. wants to thank Dr. H. Liu and Dr. J. Li of Micron Tech., Inc., Boise and Dr S. K. Banerjee and Dr. L. F. Register at the University of Texas at Austin for their help and support.

References

[1] K. Bernstein *et al.*, “High-performance CMOS variability in the 65-nm regime and beyond,” *IBM J. Res. & Dev*, vol. 50, no. 4/5, 2006.

[2] Synopsys Inc., TCAD Sentaurus, Version A-2007.12, 2007.

[3] S. Boning and P.K. Mozumder, “DOE/Opt: A system for design of experiments, response surface modeling, and optimization using process and device simulation,” *IEEE Trans. Semicond. Manuf.*, vol. 7, no. 2, pp. 233-244, May1994.

[4] C.M. Pichler, R. Plasun, R. Strasser and S. Selberherr, “Simulation of complete VLSI fabrication processes with heterogeneous simulation tools,” *IEEE Trans. Semicond. Manuf.*, vol. 12, no. 1, pp. 76-86, Feb. 1999.

[5] S. Williams and K. Varahramyan, “A new TCAD-based statistical methodology for the optimization and sensitivity analysis of semiconductor technologies,” *IEEE Trans. Semicond. Manuf.*, vol. 13, no.2, pp. 208-218, May 2000.

[6] A. A. Mutlu and M. S. Rahman, “Statistical methods for the estimation of process variation effects on circuit operation,” *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 4, pp. 364-375, Oct 2005.

[7] C. Mouli K. Prall and C. Roberts, “Trends in memory technology – reliability perspectives, challenges and opportunities,” *Proc. of the 14th International Symp. on the Phys. & Failure Analysis of ICs*, July 2007.

[8] R. H. Myers, “Response surface methodology: process and product in optimization using designed experiments,” Wiley, 1995.

[9] A. Miller, “Subset Selection in Regression,” Monographs on Statistics and Applied Probability, Second Edition, CRC Press, 2002.

[10] M. D. McKay, R. J. Beckman and W. J. Conover, “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code,” *Technometrics*, vol. 21, no. 2, pp. 239-245, May 1979.

[11] JMP, Version 7. SAS Institute Inc., Cary, NC, 1989-2007.

[12] D. C. Montgomery, “Introduction to Statistical Quality Control”, Wiley, 2004.

[13] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, “Design and Analysis of Computer Experiments,” *Statistical Science*, vol. 4, no. 4, pp. 409-435, Nov. 1989.