Accelerated Variation Simulation through Parameter Reduction

W. Paul Griffin and Kaushik Roy
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana 47907
Email: wgriffin@ecn.purdue.edu
Telephone: (765) 494–0759

***Abstract*—Proper understanding of the effects of parameter variations in a circuit requires simulation; unfortunately, accurate variation simulation can hinder simulation performance. Even though the majority of interdie variations occur between four parameters (L, W, t_{ox}, V_{fb}) [1], this parameter set is still too big for efficient large-scale simulations. As these variations all affect threshold voltage (V_{th}), ΔV_{th} is often used as a substitute for variations [2], [3]. While a ΔV_{th} substitute offers vast improvement in runtime, it has a rarely-understood loss in quality.**

In this work, we demonstrate two methods that, when presented with a set of device variations (L, W, t_{ox}, V_{fb}) and a model, can simplify those variations into a reusable model that provides accelerated simulation. Our first method, approximate ΔV_{gs} superposition, offers an accelerated method to reduce process variations down a single, manageable ΔV_{th}-like parameter.

Our second, reduced parameter method, preserves the effects of individual variations while lowering the runtime complexity. Instead of generating device attributes specific to the tested circuit, our method used a *dynamic superposition* approach to interpolate device parameters from a reduced-dimension lookup table.

Both of our methods demonstrate significant runtime computation savings, with a low break-even point as compared to the original model. While the ΔV_{gs} approach does demonstrate a noticeable quality loss compared to the original model, our reduced parameter approach demonstrates minimal loss in quality.

***Index Terms*—Parameter variations, nanoscale, MOSFET**

I. Introduction

To avoid costly manufacturing mistakes, variations' effect on circuit behavior is predicted before manufacture through the use of simulation tools. Monte Carlo analysis, the most frequently-used approach, generates a representative set via random circuit simulations.

As Monte Carlo analysis suffers from slow convergence, measures are used to accelerate the per-measurement cost of simulation. Tabular representation of a model allows for fast and reasonably accurate simulation, but is subject to a few constraints. The granularity of the tabular data can restrict the quality of interpolation; too few points and the underlying characteristics could be lost. The parameter count plays a role in both space and speed - as the number of table parameters increases, the table size and the interpolation delay both grow exponentially.

While a simple DC transistor model could be represented by a two-dimension, gate voltage and drain voltage [V_{gs}, V_{ds}] table, such a table is inadequate for variations. For interdie variations, every Monte Carlo run would require a new transistor table that could be shared between identically-sized transistors. Intradie variations would require a fresh table for every transistor in each Monte Carlo run.

A threshold voltage (V_{th}) approximation of variations can condense a model down to a simple [$V_{gs} - \Delta V_{th}$, V_{ds}] table: gate voltage V_{gs} and drain voltage V_{ds}, with gate voltage fluctuations according to the effect of variations on threshold voltage ΔV_{th}.

While the most frequent-used variation approximation is a normal ΔV_{th} distribution, others have developed approaches to efficient representation of variations. Drennan et al. [3] constructed a formula to calculate ΔV_{th} from more basic variations. One additional avenue is a $\Delta I_{on}/I_{on}$ mismatch model, but with simplifications made either through a combination of principal component/region-specific modeling [4] or Taylor series expansion [5], [6].

In this work, we present two generalized methodologies for parameter reduction. Via small-signal analysis and superposition, we form a first-order, minimial-parameter ΔV_{gs} approximation of variations. Using improved techniques, we demonstrate a second-order, "reduced parameter" method using *dynamic* superposition which offers greater accuracy during circuit analysis with only a minor reduction in efficiency.

II. Approximate ΔV_{gs} Superposition

As the most frequently-used method for variation simulation is a ΔV_{th} approximation, a similar simplification methodology via a gate voltage shift (ΔV_{gs}) definition (Fig. 1) was sought as it allows current expression via a 2D shared ($I[V_{gs} - \Delta V_{gs}, V_{ds}]$) array. In the process, an accelerated method to determine the ΔV_{gs} distribution parameters (e.g., σV_{gs} was achieved.

For two parameters to have an independent effect on the output, their output covariance ought to be zero. For any given parameter in which the relative variance (σ_x/μ_x) is small, the covariance between that parameter and others on an output is likely to be small. Furthermore, a small relative variance lets us use small-signal analysis approximations to assume that the

978-1-4673-5736-4/13/$31.00 ©2013 IEEE

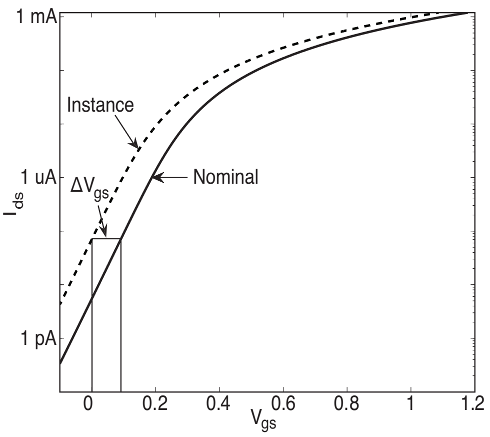


Fig. 1. Calculation of ΔV_{gs} via current matching techniques. The properties of a transistor with variations - a single "instance" - are compared against the nominal, variation-free transistor by observing the shift in gate voltage required for the nominal transistor to match the current draw of the simulated instance at the prescribed voltage (in this example, 0 V).

translation of a normal distribution to the output domain will also be normal. To find the output variance, one must first find the relationship between each parameter and the output, and then combine their effects together via a convolution of independent normal distributions.

First, the small-signal relationship between minute variations in a basic parameter (e.g., L) and an output measurement (I) is determined from a first-order Taylor series approximation.

$$I(L+\Delta L) \approx I(L) + \frac{dI}{dL}\Delta L \tag{1}$$

$$\Delta I(\Delta L) \approx \frac{dI}{dL}\Delta L \tag{2}$$

Repeating the same procedure for ΔV_{gs} and I can link together ΔL's effect on ΔV_{gs} in terms of a slope factor $m_{L\rightarrow Vgs}$.

$$\begin{aligned} \Delta I &= \frac{dI}{dL}\Delta L = \frac{dI}{dVgs}\Delta V_{gs} \\ \Delta V_{gs} &= \frac{dV_{gs}}{dI}\frac{dI}{dL}\Delta L = m_{L\rightarrow V_{gs}}\Delta L \end{aligned} \tag{3}$$

If one then assumes that the input variations are following a normal distribution and have an independent effect on the output, the output will also be normal. At a given critical point I_C, the variances are thereby additive.

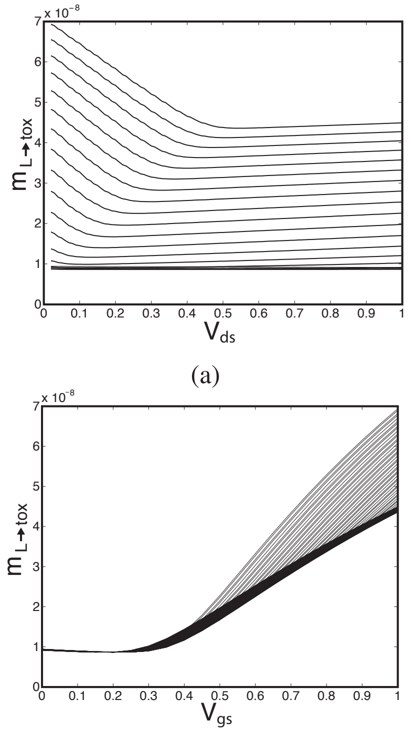
$$\begin{aligned} \sigma_{vgs}^2 &\approx \left[\frac{dV_{gs}}{dI_C}\frac{dI_C}{dL}\sigma_L\right]^2 + \left[\frac{dV_{gs}}{dI_C}\frac{dI_C}{dW}\sigma_W\right]^2 + ... \\ &\approx (m_{L\rightarrow Vgs}\sigma_L)^2 + (m_{W\rightarrow Vgs}\sigma_W)^2 + ... \end{aligned} \tag{4}$$

III. Reduced parameter via Dynamic superposition

While the small-signal approach in Section II simplifies all the variations to one parameter, it loses accuracy away from the critical point I_C. The relationship between parameters is nonlinear, and as such, ΔV_{gs} should not be normal.

TABLE I
Normal Variable ($x = \mu_x + \sigma_x z$) translations vs. Standard Normal Distribution translations.

Variable Transform	Standard Distribution Transform
$y(x) = m\ x + b$	$y(z) = \mu_y + m\ \sigma_x z$
$y(x) = m \ln(x) + b$	$y(z) = \mu_y + m \ln\left[1 + \frac{\sigma_x z}{\mu_x}\right]$
$y(x) = m\ x^{\ \ 1} + b$	$y(z) = \mu_y + m\left[(\mu_x + \sigma_x z)^{\ \ 1} + \mu_x{}^{1}\right]$
$y(x) = e^{m\ x+b}$	$y(z) = \mu_y e^{m\ \sigma_x z}$



(a)

(b)

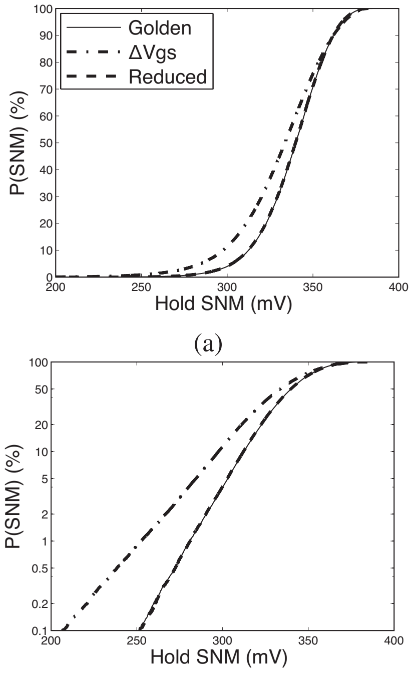
Fig. 2. Slope factors. The relationship between variations in L and t_{ox} (with respect to I) changes based on the region of operation. Shown in terms of (a) V_{ds} and (b) V_{gs} with steps in V_{gs} and V_{ds}, respectively.

The notion of independence is, however, a powerful means for combining together parameters. If one can identify sets of independent relationships, one can achieve variable reduction through superposition while still preserving covariance when it does occur.

Using Table I, one can see how some normal variable transformations are modulation-based (e.g., follow the form $y = \mu_y + f(\mu_x, \sigma_x)$) while others use scaling ($y = \mu_y \cdot f(\mu_x, \sigma_x)$). For modulated random numbers, superposition of their effects would yield a composite random number generator of the form

$$y_C = \mu_y + \sigma_y z + f_x(\mu_x, \sigma_x) + ... \tag{5}$$

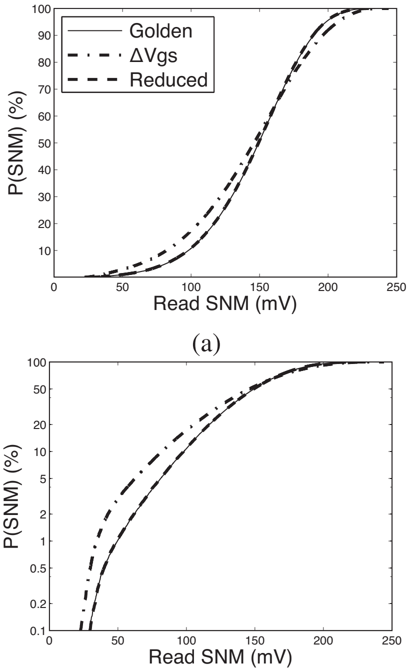
In practice, this relationship between parameters is inconsistent. While at any given point across the operation region [V_{gs}, V_{ds}] two parameters may be considered independent, the slope factors m from Table I will change. Figure 2 demonstrates such an observed relationship. As such, a *dynamic* slope factor m based on the region of operation is required.



(a)

(b)

Fig. 3. SRAM static noise margins under hold conditions, shown in (a) linear and (b) logarithmic scales.



(a)

(b)

Fig. 4. SRAM read noise margins under read conditions, shown in (a) linear and (b) logarithmic scales.

$$y_C = \mu_y + \sigma_y z + m_{x\to y}[V_{gs}, V_{ds}] \cdot f_x(\mu_x, \sigma_x) + ... \tag{6}$$

While this dynamic nature disrupts the ability to describe the composite generator by *any particular* distribution, simulations do not require the distribution shape; only its random variable generator.

Application of this approach to Pao-Sah's double integral [7] results in two sets of independent relationships: a trivial relationship between V_{fb} and ΔV_{gs} ($\sigma V_{gs} = \sigma V_{fb}$), and a composite, logarithmic-based one between W, L, and t_{ox} for three standard normal random number generators (z_{tox}, z_L, z_W).

$$\begin{aligned} t_{oxC} =& \mu_{tox} + \sigma_{tox} z_{tox} + m_{L\to tox}[V_{gs}, V_{ds}] \ln\left(1 + \frac{\sigma_L z_L}{\mu_L}\right) \\ &+ m_{W\to tox}[V_{gs}, V_{ds}] \ln\left(1 + \frac{\sigma_W z_W}{\mu_W}\right) \end{aligned} \tag{7}$$

In doing so, the model's current flow can be captured by a two-level lookup table: first via the two-dimensional slope matrices from Equation 7, and then a reduced, three-dimensional shared matrix produces the correct current value.

$$I = I_{ds,nom}[V_{gs} - \Delta V_{gs}, V_{ds}, t_{oxC}] \tag{8}$$

IV. Circuit Simulation

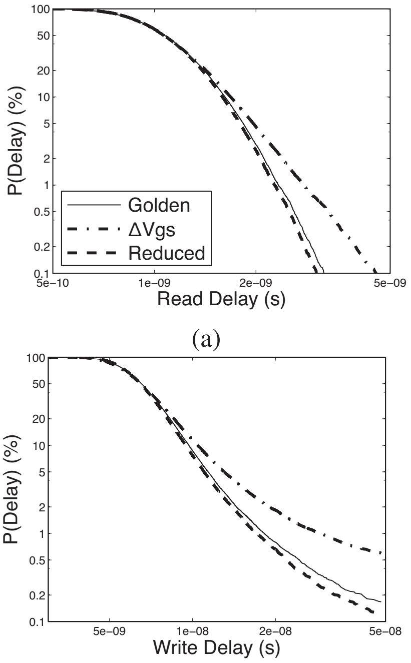
To verify these simplified modeling techniques, we constructed a tool in C to accelerate calculation of Pao-Sah's Double Integral and to provide for netlist functionality. Using a 65nm feature size, we tested our models using a 6T SRAM cell as well as a three-inverter ring oscillator.

The resource usage for several simulations are shown in Table II. While both simplifications require precomputation to generate the non-transistor-specific models, during simulation, they require little to no per-circuit setup and only induce a small performance overhead during the runtime loop.Our implementation required about 10 MB of RAM for the 2D ΔV_{gs} model, but the reduced parameter method required closer to 2 GB of ram due to the 3D lookup tables

For the SRAM cell, we tested hold static noise margin (SNM) (Fig. 3), read SNM (Fig. 4), and access delay (Fig. 5). As can be observed, the ΔV_{gs} approach consistently showed error across the SRAM attributes, while the reduced parameter approach held closely to the golden model.

The ring oscillator was tested under normal (1.0 V) and low-power (0.6 V) conditions (Fig. 6) to demonstrate the effects of a varying supply voltage. While the ΔV_{gs} parameter does show error even under nominal conditions, it is exacerbated under low-power conditions. Thanks to its point calibration approach, the ΔV_{gs} distribution holds far less accuracy in the tail, while the reduced parameter's dynamic nature consistently holds close to the golden distribution.

The tail analysis of these two representative circuits demonstrates an important note: using a ΔV_{gs} simplification can incur an important error during circuit simulation. However, with our dynamic approach to interpolation, it is possible to save processing resources while preserving quality.



(a)

(b)

Fig. 5. SRAM access delay distribution, enhanced to demonstrate right tail accuracy. (a) demonstrates the amount of time necessary to perform a successful read ($V_{bl} < 0.9\ V$) while (b) demonstrates the time necessary to perform a successful write ($V_q < 0.1\ V$; $V_{qb} > 0.9\ V$)

TABLE II
SIMULATION RUNTIME FOR 1000 CIRCUITS.

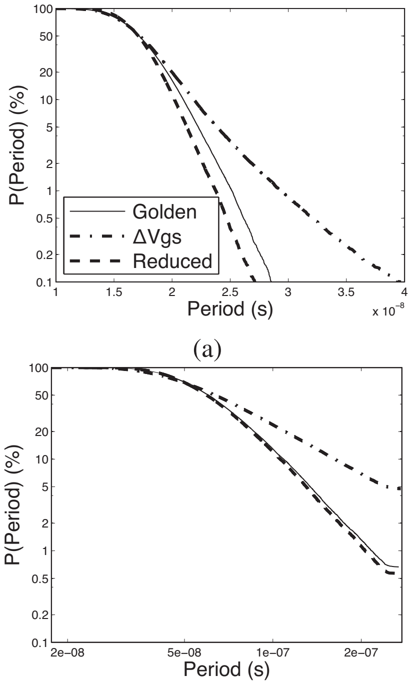
	SRAM SNM		SRAM Access		Osc. period	
	Hold	Read	Read	Write	1.0 V	0.6 V
Iterations/circuit	44.2k	33.4k	11.4k	76.2k	411k	176k
Golden Model						
Setup (s)	1358	1357	1357	1356	1367	1364
Runtime (s)	83.2	62.2	33.3	267	1474	598
ΔV_{gs} Model						
Precompute (s)	37.0	37.3	37.0	37.0	37.0	36.9
Setup (*ms*)	23.6	19.1	14.3	15.9	20.0	18.0
Runtime (s)	150	115	52.3	405	2177	1040
Circuit Speedup	9.6x	12.3x	26.6x	4.0x	1.3x	1.89x
Break-Even (*ckts.*)	28.6	28.6	27.5	30.3	55.6	40.0
Reduced Parameter Model						
Precompute (s)	47.4	47.9	47.7	47.8	47.7	48.0
Setup (*ms*)	26.4	25.9	11.7	16.2	21.8	16.4
Runtime (s)	99.9	75.9	36.6	290	1617	650.4
Circuit Speedup	14.4x	18.7x	38.0x	5.6x	1.76x	3.02x
Break-Even (*ckts.*)	35.3	35.6	35.3	35.8	39.0	36.6

V. CONCLUSIONS

In this paper we presented two methods to reduce a physical device model to allow for accelerated circuit simulation under variations.

Our first method, a ΔV_{gs} approach, demonstrates both a ΔV_{th} approach to accelerated simulation, as well as an accelerated way to construct such a model without the use of Monte Carlo simulations. However, this same approach demonstrates a critical fact: threshold variation models are inadequate. They can only be considered accurate for a single operating point dependent on the calibration criteria.

Our second approach to parameter reduction demonstrates how to apply superposition without a loss in quality. Through recognition of algebraically-consistent relationships between variables and allowance for varying coefficients, we can reduce the number of parameters required for our simplified model. This intelligent superposition approach allows for increased accuracy, providing us with a level of accuracy far exceeding a ΔV_{th} approximation.



(a)

(b)

Fig. 6. Ring oscillator period under the presence of variations. (a) demonstrates the ring oscillator period under standard conditions (Vdd=1 V), while (b) demonstrates the reduced accuracy achieved under low-power (Vdd=0.6 V) conditions.

ACKNOWLEDGMENTS

This research was funded in part by the National Science Foundation.

REFERENCES

[1] B. Ping Yang and P. Chatterjee, "Statistical modelling of small geometry MOSFETs," in *IEEE IEDM*, vol. 28, 1982, pp. 286–289.

[2] T. Tsunomura, A. Nishida, and T. Hiramoto, "Verification of threshold voltage variation of scaled transistors with ultralarge-scale device matrix array test element group," *Japanese Journal of Applied Physics*, pp. 124 505–1 – 124 505–4, 2009.

[3] P. Drennan and C. McAndrew, "A comprehensive MOSFET mismatch model," in *IEEE IEDM*, 1999, pp. 167 –170.

[4] V. Wang, K. Agarwal, S. Nassif, K. Nowka, and D. Markovic, "A simplified design model for random process variability," *IEEE TSM*, pp. 12 –21, Feb 2009.

[5] J. Croon, M. Rosmeulen, S. Decoutere, W. Sansen, and H. Maes, "An easy-to-use mismatch model for the mos transistor," *IEEE JSSC*, pp. 1056 – 1064, Aug 2002.

[6] Q. Zhang, J. Liou, J. McMacken, J. Thomson, and P. Layman, "Spice modeling and quick estimation of mosfet mismatch based on bsim3 model and parametric tests," *IEEE JSSC*, pp. 1592 –1595, Oct 2001.

[7] C. Sah and H. Pao, "The effects of fixed bulk charge on the characteristics of metal-oxide-semiconductor transistors," *IEEE Transactions on Electron Devices*, pp. 393 – 409, Apr 1966.