Fast Statistical Circuit Analysis with Finite-Point Based Transistor Model

Min Chen Wei Zhao Frank Liu* Yu Cao
Arizona State University, Phoenix, USA, {min.chen, wei.zhao, ycao}@asu.edu
*IBM Research Laboratory, Austin, USA, frankliu@us.ibm.com

Abstract – A new approach of transistor modeling is developed for fast statistical circuit simulation in the presence of variations. For both I-V and C-V characteristics of a transistor, finite data points are identified by their physical meaning; the impact of process and design variations is embedded into these points as closed-form expressions. Then, the entire I-V and C-V are extrapolated using polynomial formulas. This novel approach significantly enhances the simulation speed with sufficient accuracy. The model is implemented in Verilog-A at 65nm node. Compared to simulations with the BSIM model, the computation time can be reduced by 7x in transient analysis and 9x in Monte-Carlo simulations.

1. Introduction

CMOS technology will arguably be the technology of choice into sub-45nm regime. Such extremely small devices inevitably suffer from severe variations due to the fabrication process and circuit operations [1-3]. Due to the statistical nature of these variations, current deterministic design flow needs to be shifted toward statistical circuit analysis and optimization. This avoids overly pessimistic design margins [4-6].

Conducting statistical circuit analysis demands highly efficient simulations. The problem of computation cost is further exacerbated as the total number of transistors in a chip rapidly increases into the gigascale regime. For instance, it usually takes more than weeks for a contemporary design to run full-chip SPICE simulations; the cost to run Monte-Carlo simulations will be inhibitive in this situation. Therefore, to enable a smooth transition to statistical design flow, it is critical to develop simulation techniques that are highly efficient in computation and keep the fidelity to parameter variations.

The speed of circuit-level simulations depends on both the complexity of the transistor model (e.g., BSIM), as well as the efficiency of circuit simulation tools (e.g., SPICE). In this work, our focus is to significantly simplify the transistor model, without sacrificing its accuracy and the sensitivity to variations. Traditional compact transistor models, such as BSIM or PSP, consist of a large number of parameters and complicated equations in order to capture many physical mechanisms for a short-channel device [7-8]. Such a high-level of complexity provides sufficient physicality of models, but dramatically slows down the simulation speed. At the other extreme end, lookup tables for I-V and C-V are purely based on the empirical behavior of a transistor [9]. It is much more efficient than the compact models, if the size of the table is small. However, the empirical nature of lookup tables limits their scalability to process variations. As the number of process corners increases, the size of lookup tables rapidly rises and thus, leads to higher simulation cost. In this context, a desirable approach should be to have the efficiency comparable to lookup tables, but keep the scalability of compact models.

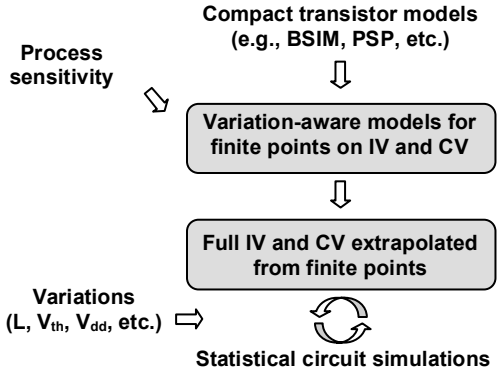

Figure 1. Application of new model in the design flow

To achieve this objective, a finite-point based transistor model is proposed in this paper. The key idea is to extract only a finite number of data points from I-V, based on their physical meanings and the importance in circuit operation. The dependence on process and design variations is embedded in these points through physical models or lookup tables. The entire behavior of I-V and C-V is then extrapolated from these points with simple polynomial equations (Section 2). Such an approach combines the simplicity of lookup tables and physically maintains the dependence on variations. Fig. 1 illustrates the application of this model in statistical design flow to replace traditional compact models. Its accuracy is calibrated by compact models, while this approach is general for all types of compact models and circuit simulations. Section 3 describes the implementation of the model in Verilog-A and evaluates the convergence, accuracy, and speed as compared to BSIM4 at 65 nm node. Finally, the work is concluded in Section 4.

978-3-9810801-2-4/DATE07 © 2007 EDAA

2. Finite-Point Based Transistor Model

The I-V and the C-V characteristics of a transistor determine its behavior in any circuit operation. They are modeled in Section 2.1 and 2.2, respectively. Their variability due to variations in the threshold voltage (V_{th}) and the effective channel length (L_{eff}) are calibrated by the variational model discussed in Section 2.3.

2.1 I-V modeling

The I-V characteristic of a transistor can be classified into three regions: saturation, linear, and subthreshold. Thus, there are three points that are the most important to discriminate these regions. We use X($V_{x,g}$, $V_{x,d}$, I_x) to represent a data point X. $V_{x,g}$, $V_{x,d}$ and I_x are gate voltage, drain voltage and the drain source current of X, respectively. As shown in Fig. 2, these three points are:

A(V_{DD}, V_{DD}, I_{on}): the maximum current I_{on}.

B(V_{DD}, V_{lin}, I_{lin}): the separation between the saturation and linear region.

C(V_{start}, V_{DD}, I_{start}): the boundary of the saturation and subthreshold region.

Given a specific technology, these points can be quantified either from compact transistor models or directly from the measurement. At 65nm node, V_{lin} of B is approximately half V_{dd}; V_{start} of C is around V_{th} + 0.1 to ensure the linearity in the saturation region.

These three points are essential to determine the entire I-V and separate different regions. From these three points, we can rely on the I-V characteristic of each region to extrapolate other points. Such extrapolations can be either physical or empirical, as long as they are sufficiently accurate.

In the saturation region, we can use a linear curve to model the dependence of I_{ds} on gate voltage V_{gs}., while I_{ds} is insensitive to V_{ds}. The parameter α in the alpha-power law is almost equal to one in the nanoscale regime [10].:

$$I_{ds} \propto W \cdot (V_{gs} - V_{th})^{\alpha} = W \cdot (V_{gs} - V_{th}) \quad (1)$$

This linear dependence is due to the strong velocity saturation of a short channel device, and is more pronounced in 90nm and 65nm nodes. In the linear region, the drive current has a quadratic dependence on V_{ds}. The choice of a quadratic function in the linear region further guarantees the first order continuity of the current from the linear to saturation region.

Besides these three primary points, two more points, D(V_{start}, V_{ssa}, I_{ssa}) and E(V_{sub}, V_{DD}, I_{sub}), are added to improve the model accuracy. As shown in Fig. 2, D improves the accuracy in the linear region, especially when V_{gs} is low. E is in the subthreshold region and defines the leakage when $V_{gs} < V_{sub}$. The leakage is further matched by fitting the subthreshold swing.

From the perspective of circuit analysis, the importance of these sampling points can be understood by studying the transistor switching trajectory [10], where the transistor is modeled as a current source driving a load capacitor. These points could be treated as input metrics in gate delay analysis instead of I_{vdsat} along to achieve better accuracy [11]. Fig. 3 shows the typical NMOS trajectories of an inverter and an NAND2 gate. In a fast input switching, the trajectory will start from point C, travel towards A and then fall to zero through B. In a slow input case, it will start from C, travel towards B then fall to zero. Therefore, by carefully matching of those critical points, we can preserve the accuracy of circuit analysis under various operation conditions. The worst case of this new model happens when there is extremely large fan-out and slow input. Fortunately, this situation is quite rare in an appropriately optimized digital design.

The closed-form piece-wise I-V model is developed based on the above definitions. The model format is summarized as equations (2)-(5):

Deep-subthreshold region:

$$I_{ds} \propto \exp\left(\frac{V_{gs} - V_{sub}}{nv_T}\right) \cdot V_{ds} \quad (2)$$

Subthreshold region:

$$I_{ds} \propto \exp\left(\frac{\ln(I_{start}) - \ln(I_{sub})}{V_{start} - V_{sub}}(V_{gs} - V_{sub})\right) \cdot V_{ds} \quad (3)$$

Linear region:

$$I_{ds} = J \cdot \frac{I_{on} - I_{lin}}{V_{DD} - V_{lin}} \cdot V_{ds} + K \cdot (V_{gs} - V_{start}) \cdot \left((M \cdot (V_{lin} - V_{ds})^2 + I_{lin}\right) \quad (4)$$

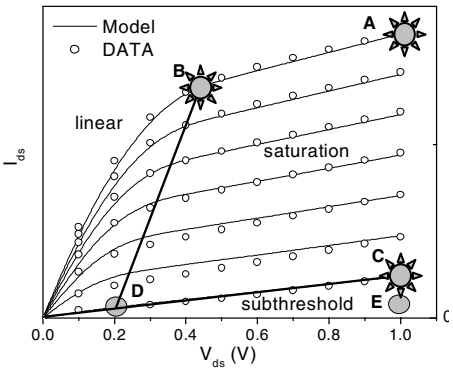

Figure 2. Critical points and model definition

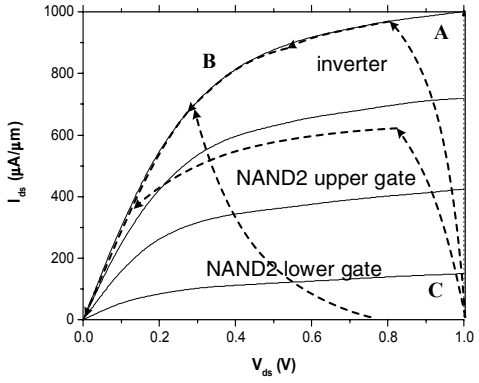

Figure 3. The switching trajectory of Inverter and NAND2

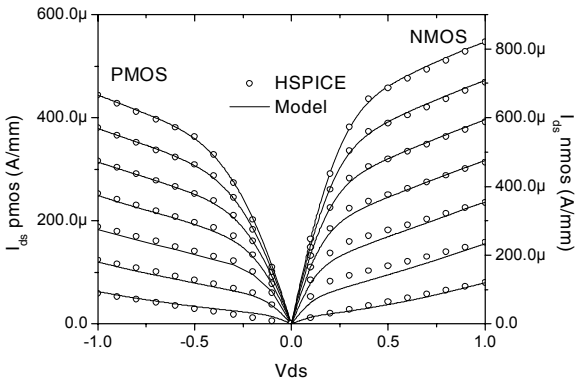

Figure 4. I-V model matching evaluation for 65nm node

Saturation region:

$$I_{ds} = J \cdot \frac{I_{on} - I_{lin}}{V_{DD} - V_{lin}} \cdot V_{ds} + K \cdot (V_{gs} - V_{start}) \cdot I_{lin} \tag{5}$$

where n is subthreshold swing and v_t is the thermal voltage. J, K, M are coefficients decided by the linear dependence of equation (1). The equations (4) and (5) have the same first term, which keeps the equation first order derivative continuous at the boundary. In turn, this guarantees the simulation convergence.

Fig. 4 shows the I-V matching of the model at the 65nm node with zero body bias and 1V supply voltage, for both PMOS and NMOS devices. When there is non-zero body bias, the model captures its impact through the parameter of V_{th} at the critical points. More details will be discussed in section 3.3 to model this effect.

2.2 C-V modeling

The transistor capacitance model can be divided into two groups, the intrinsic and the extrinsic. The intrinsic capacitances are decided by channel charges and operating regions, which are functions of both geometry and voltages. The extrinsic capacitances are decided by the dimensional parameters of the transistor (i.e., area and periphery of source and drain). In this work, the capacitance model is developed based on the derivations of BSIM4 capmod=1 [12]. The key idea is to find the charge of each node and make sure of their first-order continuity at the boundaries with minimal computation overhead.

A considerable portion of C-V equations is smoothing functions, used to enhance convergence at the boundaries of different regions. To lower the computational cost for the C-V model, we can simplify those smoothing functions to reduce computational costly terms such as square, exponential and logarithm. For example, one smoothing function used in BSIM is:

$$z = x - 0.5\left\{(x - y - \delta) + \sqrt{(x - y - \delta)^2 + 4\delta x}\right\} \tag{6}$$

which is used to connect

$$z = \begin{Bmatrix} x & x < y \\ y & x \geq y \end{Bmatrix} \tag{7}$$

It can be simplified as a piece-wise function:

$$z = \begin{Bmatrix} x & x \leq y - \sigma \\ y - (x - y - \sigma)^2 / (4\sigma) & y - \sigma < x < y + \sigma \\ y & x \geq y + \sigma \end{Bmatrix} \tag{8}$$

where σ is a fitting parameter similar to δ.

Experiments show that around 10% speed improvement can be achieved for C-V model once this second order polynomial smooth function is applied. We apply similar techniques all over the capacitance model to enhance the simulation efficiency.

2.3 Process variations

Process variations from V_{th} and L_{eff} can dramatically change the drive current of a transistor. This effect severely impacts circuit performance such as delay and leakage. Their impact is incorporated at these critical points through analytical equations. Once the variability of these critical points is decided, the whole I-V can be extrapolated using the same method in section 2.1.

In saturation and linear regions, the whole I-V is scalable with the current at A and B. Therefore, I_{on} and I_{lin} variability models are important to determine other current points. In a nanoscale transistor, short channel effects, such as drain induced barrier lowering (DIBL), are critical. In particular, V_{th} exhibits a strong dependence on L_{eff} and V_{ds} due to DIBL. Based on the physical derivation in [12], we simplify its model for DIBL as:

$$V_{th}(L_{eff}) = V_{th0} - \alpha \cdot V_{ds} \cdot \frac{0.5}{\cosh(\beta \cdot L_{eff}) - 1} \tag{9}$$

where V_{th0} is the long channel V_{th} and α, β is the DIBL coefficient. Combined with the variation of V_{th0} due to random doping, we have the close form equation:

$$V_{th} = V_{th,nominal} + \Delta V_{th}(L_{eff}) + \Delta V_{th}(V_{th0}) \tag{10}$$

in which $V_{th,nominal}$ is the threshold voltage in nominal case. Furthermore, considering the velocity saturation (I_{on} and I_{lin} both satisfies this condition), we express I_{ds} as:

$$I_{ds} \propto v_{sat} \cdot W \cdot (V_{ds} - V_{th} - V_{dsat}/2) \tag{11}$$

where v_{sat} and V_{dsat} are the velocity and voltage at saturation, respectively. Equations (9)-(11) accurately predict the current variability and apply to both I_{on} and I_{lin}.

Fig. 5 demonstrates the accuracy of this model using I_{on} of NMOS as an example in 65nm node. In the presence of both V_{th} and L_{eff} variations around ±25%, error is less than 2.5% for each of the tested corners.

The situations at point C and E are different. Taking C as an example, if the voltage V_{start} is fixed and the current I_{start} is changed, C may fall into the subthreshold region due to the reduction of current, which is not allowed. Therefore, it is more practical to fix the current

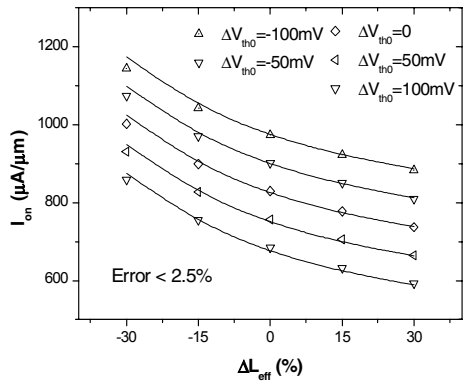

Figure 5. I_{on} variability of process variations

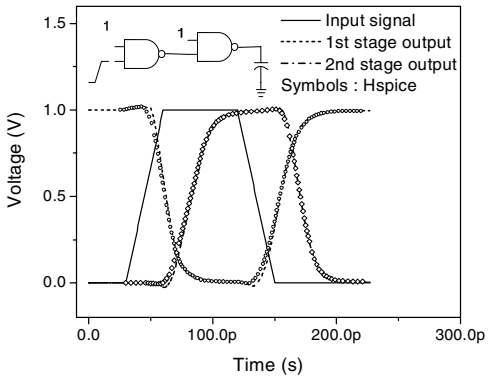

Figure 6. Waveform evaluation using 2-stage NAND2

and make voltages scalable to the variation for these two points. Using equation (9) and (10), we derive V_{start} and V_{sub} as:

$$V_{start} = V_{start,nominal} + \Delta V_{th} \tag{12}$$

$$V_{sub} = \frac{n_{nominal}}{n} \cdot (V_{sub,nominal} + \Delta V_{th}) \tag{13}$$

where subthreshold swing n could be fixed to the nominal value for the simplicity reason.

For point D, we observe the dependence of V_{ssa} *on* L_{eff}, where it can be modeled as a piece-wise linear function:

$$V_{ssa} = \begin{cases} \delta 1 \cdot \Delta L_{eff} + V_{ssa,nominal} & \Delta L_{eff} > 0 \\ \delta 2 \cdot \Delta L_{eff} + V_{ssa,nominal} & \Delta L_{eff} \le 0 \end{cases} \tag{14}$$

in which δ1 and δ2 are fitting parameters decided in the calibration.

Equations (11)-(14) propose a complete set of analytical variation models for the finite critical points. For the C-V model, because the node charges are function of V_{th} and L_{eff}, the variability due to them has already been included. With the increasing number of process corner, the finite-point model requires much smaller sizes compared to that of the lookup tables. It also benefits from excellent model scalability.

3. Model Evaluation

The proposed finite-point based model is simple in mathematics, requires a much fewer number of parameters, and ease the extraction and simulation. The evaluation of its convergence, accuracy, sensitivity and speed is presented in this section, as compared to BSIM.

3.1 Experimental setup

The model is implemented and tested under various conditions by varying input slew T_r, load capacitor C_{load}, fan-out FO, V_{th0} and L_{eff}. Low to high switching delay t_{LH}, high to low switching delay t_{HL}, 10%-90% output rising transition time t_{rise} , 90%-10% output falling transition time t_{fall}, and oscillating frequency f are set as evaluating objects. As shown in Fig. 1, the evaluation procedure is listed as following:

1) Implement the model using Verilog-A.
2) Extract five critical points with HSPICE using BSIM4 65nm PTM model card [13].
3) Calibrate the variation model for these points.
4) Simulate benchmark circuits such as inverter, NAND2, AOI and NAND chain oscillator in nominal and variational cases in SPICE.
5) Implement BSIM4 model using Verilog-A. Compare both accuracy and efficiency.

3.2 Nominal circuit analysis

The convergence and accuracy of our model can be evaluated through the switching waveform, delay, and output slew of benchmark circuits.

The switching waveforms are used to evaluate the overall quality of the model. Non-smooth waveform or even possible failure of the simulation may happen when there is any convergence problem. The new model exhibits good convergence during simulation due to the continuity of derivatives for both the I-V and C-V. As an example, the waveforms produced by a 2-stage NAND2 are evaluated in Fig. 6, in which the model convergence can be evaluated with stack gates and gate capacitance. The switching waveforms coordinate in all the switching moments and capture the overshoot/undershoot smoothly without showing any convergence difficulty. An excellent waveform matching is achieved.

The input slew and fan-out play important roles during timing analysis. To ensure the generality needed in real design, the model needs to be accurate enough in a reasonable range for both the variables. Delay and output slews of a NAND2 gate are verified in Fig. 7 and Fig. 8, in which we tune the range of input slew and fan-out respectively. Fig.7 shows good matching when input slew varies from 20ps to 250ps (FO=4). The error is less than 6%. Excellent matching is achieved in Fig. 8 when FO varies from 1 to 9 and the input slew is 60ps. The error is less than 5%. These results confirm the model accuracy and generality.

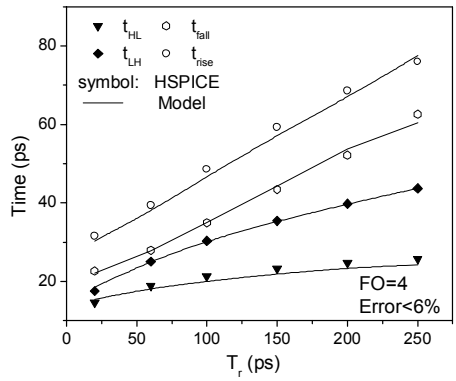

Figure 7. Output slew and delay of NAND2 with various T_r.

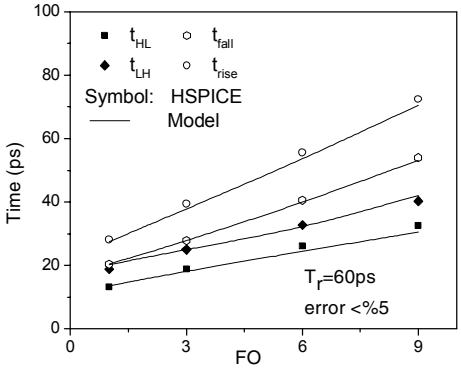

Figure 8. Output slew and delay of NAND2 with various FO.

3.3 Variational circuit simulation

The model is developed with the ability to capture delay variability under process variations. In Fig. 9, sensitivity tests are performed in NAND2 with FO = 3. Only NMOS variations are included for the simulation simplicity. As the drive current through NMOS is significantly impacted, t_{HL} suffers from a large amount of variability. Excellent matching is achieved compared with the BSIM simulation results, with error less than 6%.

To further illustrate the sensitivity of the model, the frequency of a 9-stage FO = 3 NAND2 chain oscillator is tested under a various corners. We assume that all the NMOS are 100% correlated; all the PMOS transistors are strongly correlated, but totally independent on NMOS variations. The fast (F) corner is when both V_{th} and L_{eff} have -30% variations; the slow corner (S) has 30% variations. Results are listed in Table 1. The overall error is less than 5% in this study. It demonstrates the model's capability to match all the extreme cases under process variations.

In statistical circuit simulation, the model is required to predict not only the nominal, but more importantly, the distribution of gate delay under statistical variations. We implement the model into Monte-Carlo simulation to get the probability density function (PDF) of delay. Based on

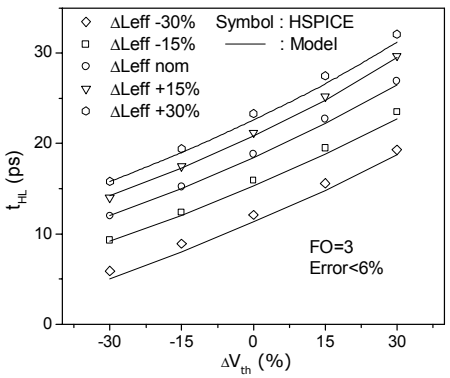

Figure 9. T_{HL} of NAND2 under NMOS variations

Table 1. Comparison of frequency *f* of 9-stage NAND2 ring oscillator (FO=3).

Variation Corners		Frequency (Ghz)		
NMOS	PMOS	Model	HSPICE	Error
F	F	5.38	5.52	2.5%
F	N	4.12	4.31	4.4%
F	S	3.15	3.15	0%
N	F	3.85	3.98	0.7%
N	N	2.89	2.98	2.7%
N	S	2.27	2.21	2.7%
S	F	2.82	2.84	0.7%
S	N	2.11	2.08	1.4%
S	S	1.66	1.59	4.4%

the assumption that both NMOS and PMOS have independent L_{eff} and V_{th} variations that follow the Gaussian distribution with 3σ at 30% of mean and 30mV respectively, 1000 times of SPICE simulations are performed for statistical analysis. Fig. 10 shows an example of the NAND2 gate delay distribution under various input slews. With larger input slew, the distribution spreads to a wider range. The model successfully captures this trend. A more complicated AOI gate is further implemented to check the delay distribution under two different switching conditions, i.e., the worst case and the best case. The worst case is when NAND2 switches with the lower gate and NOR2 switches with the upper gate; vise versa is the best case. As shown in Fig. 11, the worst case switching leads to a wider delay distribution, since the variations of both stack transistors affect the delay. Our model accurately captures this phenomenon with error of less than 6%.

3.4 Simulation speed

In addition to sufficient accuracy, the newly developed model achieves significant speed-up from the model simplicity. Fewer parameters and lower order polynomial equations greatly improve the simulation efficiency. Speed tests are conducted in both nominal transient simulations and Monte-Carlo simulations using various benchmark circuits in SPICE. The CPU running

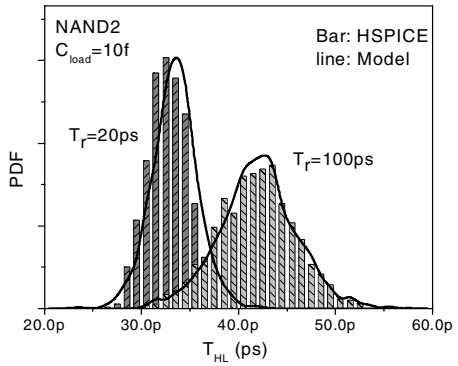

Figure 10. The distribution of delay variability of NAND2

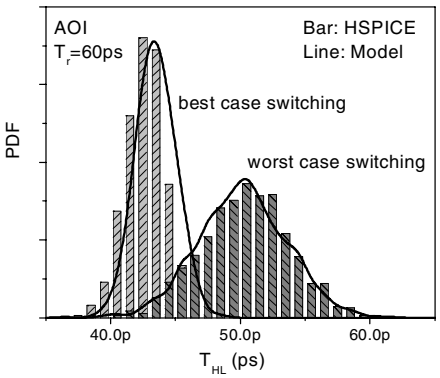

Figure 11. The distribution of delay variability of AOI

time are compared between our model and the BSIM4 model. To make it generic, two step size modes, fixed step size and dynamic step size, are used in the tests. The result from the first mode clearly indicates the computation speed of the model, while the result from the second mode reflects the iteration condition of the simulator. Test results are listed in Tab. 2, which shows that our model is at least six times faster than BSIM4 in fixed step size mode (maximal step size <1ps) and even faster in dynamic st ep size mode due to less iterations. Monte-Carlo simulation results in Table 2 show a further speed-up. The speed of statistical simulations is nine times faster due to significant reduction in setup time, which is the result of

Table 2. Model simulation speed compared with BSIM4.

Transient Simulation						
	Transient time (10ns)	BSIM4		Finite Point		Speed
		step	time(s)	step	time(s)	
9-stage FO3 NAND2 oscillator	Max step=1ps	10008	314	10073	48	**1/6.5x**
	Dynamic step	4310	156	2993	19	**1/8.2x**
4-bit static adder	Max step=1ps	10289	587	10190	95	**1/6.2x**
	Dynamic step	4232	293	3583	43	**1/6.8x**
Monte-Carlo Simulation (1000 times)						
	Transient time (1000*10ns)	BSIM4 Total time		Finite Point Total time		Speed
AOI	Dynamic step	556 (s)		58 (s)		**1/9.6x**
NAND2 oscillator	Dynamic step	1810 (m)		195 (m)		**1/9.3x**

much fewer parameters. This advantage of shorter setup time is especially important when the total number of transistors keeps increasing in a VLSI design. The speed test results along with other evaluation results indicate that our model improves model efficiency significantly without any tradeoff of accuracy.

4. Conclusion

The innovation of this work is to develop a transistor model based on five critical data points. The impact of process variations is further embedded through these points. Using simple equations to extrapolate I-V and C-V, the model achieves high speed with excellent accuracy in statistical circuit analysis. Experimental results prove the simulation efficiency over conventional compact transistor models. This model is compatible with SPICE simulators. It can be easily calibrated by various compact models, such as BSIM or PSP, or silicon measurement.

Acknowledgment

The authors would like to thank N. Hakim at Intel, K. Singhal and M. Shahram at Synopsys for valuable discussions to this project. This work is partially sponsored by MARCO MSD and C2S2, and Intel.

References

[1] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE TED*, vol. 45, no. 9, pp. 1960-1971, 1998.

[2] D. Boning and S. Nassif, "Models of process variations in device and interconnect," *Design of High-Performance Microprocessor Circuit*, Chapter 6, pp. 98-115, IEEE Press, 2000.

[3] M. Orshansky, K. Milor, P. Chen, K. Keutzer, and C. Hu, "Impact of spatial intrachip gate length variability on the performance of high-speed digital circuits," *TCAD*, vol. 21, pp. 544-553, 2002.

[4] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst case timing analysis," *DAC*, pp. 556-561, Jun. 2002

[5] C. Viswesvariah, "Statistical timing of digital integrated circuits," *ISSCC*, 2004.

[6] Y. Cao and L. Clark, "Mapping statistical process variations toward circuit performance variability: An Analytical Modeling Approach," *DAC*, pp. 658-663, 2005.

[7] Y. Cheng, et al., "A physical and Scalable *I-V* Model in BSIM3v3 for Analog /Digital Circuit Simulation," *IEEE Trans. Electron Devices*, vol. 44, no. 2, pp. 277-287, Feb. 1997.

[8] G. Gildenblat, et al., "PSP: an advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Trans. Electron Devices*, vol. 53, no. 9, pp. 1979-1993, Sep. 2006,

[9] D. Nadezhin, et al., "SOI transistor model for fast transient simulation," *ICCAD*, pp 120-127, Nov. 2003.

[10] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," JSSC, vol. 25, no. 2, pp. 584-594, Apr. 1990.

[11] M. H. Na, E.J. Nowak, W.Haensch, and J. Cai, "The effective drive current in CMOS Inverters," *IEDM*, pp. 121-124, 2004

[12] BSIM4 MOSFET Model – User's Manual, 2006.

[13] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm early design exploration," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2816-2823, Nov. 2006.