

# Super Fast Physics-Based Methodology for Accurate Memory Yield Prediction

Rajiv V. Joshi, *Fellow, IEEE*, Keunwoo Kim, *Senior Member, IEEE*, Rouwaida Kanj, *Senior Member, IEEE*, Ajay N. Bhoj, Matthew M. Ziegler, Phil Oldiges, *Senior Member, IEEE*, Pranita Kerber, Robert Wong, Terence Hook, Sudesh Saroop, Carl Radens, and Chun-Chen (Frank) Yeh

**Abstract**—We propose an efficient physics-based mixed-mode statistical simulation methodology for nanoscale devices and circuits. Here, 3-D Technology Computer Aided Design models pose a barrier for efficient simulation of variability as they generally involve millions of nodes in their mesh representations. The proposed methodology, which has been implemented for FinFET/tri-gate static random access memory (SRAM) design, overcomes this barrier by leveraging advanced physics-based 2-D (P2-D) devices with optimized meshes that are derived from 3-D FinFET models with tuned device parasitics. This enables physics-based simulation as well as physics-based variability input parameters. To improve accuracy, an embedded automated flow enables extraction of all external nodal parasitics, directly from a 3-D FinFET circuit layout representation. The circuits consisting of advanced P2-D devices are then back annotated with the nodal parasitics to enable fast and accurate SRAM dynamic margin mixed-mode simulations. Results demonstrate up to 200× speedup compared with traditional 3-D device simulations, and around five orders of magnitude wall clock time improvement on account of fast statistical methodologies, which are superior in comparison with traditional Monte Carlo analysis. This makes it feasible to supplant often inaccurate compact model-based simulations by true mixed-mode device simulations in statistical engines. The proposed physics-based methodology is also shown to corroborate well with hardware measurements.

**Index Terms**—Capacitance, fast statistical sampling, FinFET, physics-based models, static noise margin (SNM), static random access memory (SRAM), Technology Computer Aided Design (TCAD).

## I. INTRODUCTION

CHIP design with iterative fabrication cycles is often very expensive and very slow. Fabrication steps involve processes like lithography, doping, etching, chemical mechanical polishing, and hence forth. Manufacturing processes, however, are not deterministic, and have inherent

Manuscript received June 22, 2013; revised December 24, 2013; accepted March 11, 2014. Date of publication April 18, 2014; date of current version February 20, 2015.

R. V. Joshi, K. Kim, A. Bhoj, M. Ziegler, P. Oldiges, P. Kerber, C. Radens, and C. C. Yeh are with the IBM TJ Watson Laboratories, Yorktown Heights, NY 10598 USA (e-mail: rvjoshi@us.ibm.com; kkim2@us.ibm.com; ajay.bhoj@gmail.com; zieglerm@us.ibm.com; poldiges@us.ibm.com; k\_pranita@us.ibm.com; radens@us.ibm.com; yehc@us.ibm.com).

R. Kanj is with the American University of Beirut, Beirut 1107 2020, Lebanon (e-mail: rouwaida.kanj@gmail.com).

R. Wong, T. Hook, and S. Saroop are with IBM Technology, Hopewell Junction, NY 12533 USA (e-mail: rwong@us.ibm.com; thook@us.ibm.com; ssaroop@us.ibm.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TVLSI.2014.2313815

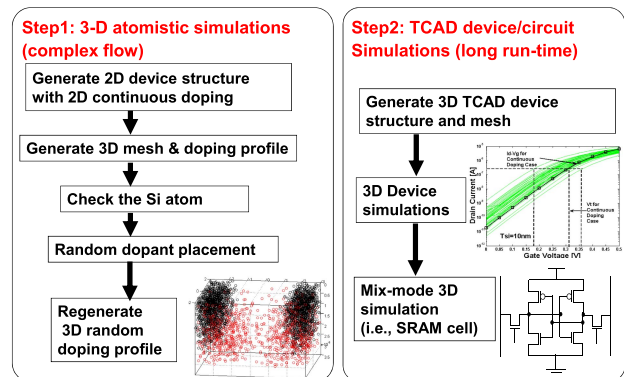


Fig. 1. General TCAD flow for circuit analysis: complex flow with long run-times. Flow is not directly applicable to DfM.

random and systematic variability built into them. Design-for-manufacturing (DfM) methods can predict or project fabricated results using software-based analysis tools. Particularly, several statistical design methodologies have been developed to address the memory design yield degradation problem. To speed up the analysis, compared with traditional Monte Carlo, fast statistical sampling methods often rely on variance reduction methods [1]–[4]. The projected results are, however, only as good as the underlying models. This poses a challenge in both present and future technology generations where fast and accurate circuit simulations are required to capture rapidly changing device features and intrinsic device fluctuations. This makes it extremely difficult for compact models (CMs) to catch up with process/technology changes.

Technology Computer Aided Design (TCAD) tools can be one of the key components of DfM. However, numerical device simulations [5] are based on solving drift-diffusion or hydrodynamic transport equations, which are cumbersome to solve. Furthermore, device/circuit mixed-mode simulations consume additional effort and more time, so TCAD tools cannot be directly used in circuit analysis. TCAD-embedded statistical analysis is intractable in the DfM world due to unacceptably high computational time, complexity of atomistic variability modeling, and weak TCAD for manufacturability (TfM) infrastructure, as shown in Fig. 1. Overall, TCAD simulations cannot be directly inserted into current generation DfM flows.

In this paper, we propose a methodology for improved circuit design and manufacturability by providing a method for TfM of process-sensitive circuits. Based on physical and process variability sources, a virtual representation of the

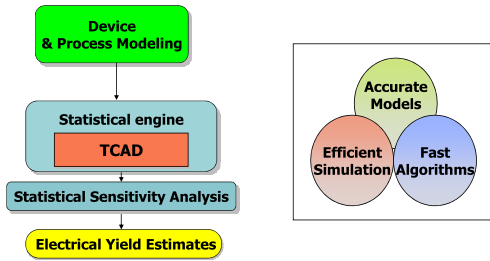


Fig. 2. Newly proposed integrated Tfm flow.

fabricated system is obtained using physical TCAD parameters that can account for process variations during production. The core idea of the proposed flow is to build a robust and fast Tfm framework by interfacing TCAD and fast statistical sampling techniques [1]–[4], and to convert device variability sources to TCAD parameters for fast statistical analysis. The proposed method is much more accurate and efficient to characterize and predict yield, and can be generalized to a variety of applications. Most importantly, the approach is pragmatic on account of the following key features: 1) optimized physics-based 2-D (P2-D) mesh representations of the 3-D devices; 2) advanced automated circuit parasitic extraction for generic FinFET circuits; 3) efficient TCAD parametric representation [e.g., gate work function (WF)] for device variability including random dopant fluctuation (RDF); and 4) a fast statistical sampling-based simulation engine. Once the hardware data are available, accuracy can be improved through hardware corroboration. Those features in turn enable accurate device models, efficient TCAD simulations, and fast statistical algorithms that constitute the pillars of a robust Tfm methodology.

II. PROPOSED METHODOLOGY

Fig. 3 shows an overview of the proposed methodology flow for Tfm of a process-sensitive circuit. It includes any circuit whose function and quality can be affected by the circuit fabrication process. Exemplary process-sensitive circuits include but are not limited to memory cell arrays, such as static random access memory (SRAM) and dynamic RAM cells. Namely, both the functionality and performance of memory cells, can be affected by process variations, such as RDFs, which result in threshold voltage ( $V_t$ ) variations, and patterning/lithographic process variations. As shown in Fig. 3(a), variability sources can include gate length (L), channel width (W), RDF, line edge roughness, oxide thickness (Tox), parasitic and/or distributed capacitance and/or resistance, supply voltage ( $V_{dd}$ ), cell-supply voltage (Vcs), chip temperature, design parameters, and aging effects like negative bias temperature instability, positive BTI, and hot carriers injection. In this paper, we focus on capturing RDF effects in an effort to demonstrate how the Tfm concept can be implemented efficiently.

As shown in Fig. 3(b), a TCAD device structure can be generated *in lieu* of many mesh structures depending on the variability sources. Then, TCAD input parameter effects can be physically evaluated based on model equations and the underlying device structure. There is no need for device

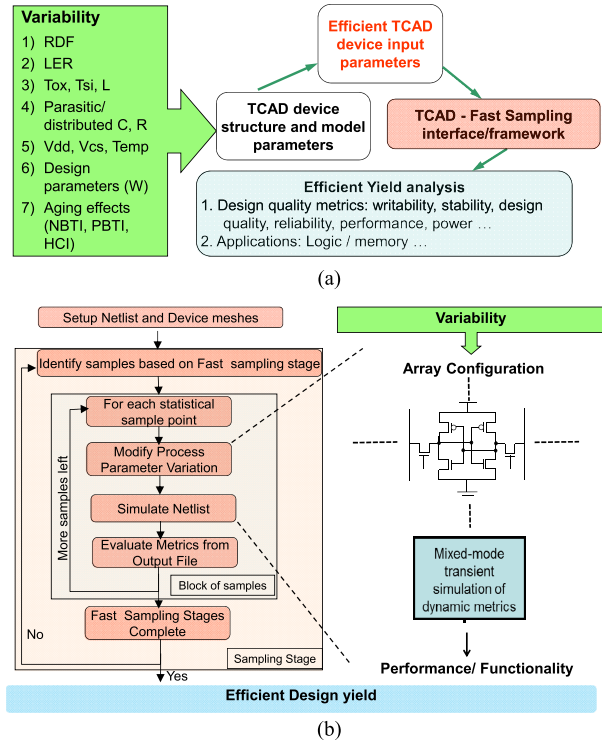


Fig. 3. (a) Proposed Tfm methods. (b) Corresponding yield analysis flow uses one TCAD structure in lieu of many meshes. Example is for an SRAM cell.

mesh regeneration with every sample point. The parameters hence can be directly used in statistical simulations. Each simulation depicts a sample point evaluation, and a collection of sample points is analyzed to provide statistical inference of a design quality metric. This flow can be applied to the analysis of memory cell performance/power/stability and other yield metrics.

- The combined interface brings forth many advantages.
- 1) The fast statistical analysis enables estimating low fail probability with a reasonably low number of samples involved. This makes it more practical to perform statistical analysis at the mixed-mode level, and allows the usage of TCAD-based statistical design and yield analysis.
  - 2) The TCAD mixed-mode analysis enables more physical and accurate analysis of the effects of the statistical process parameter variation on the yield. Such effects may not be well abstracted or captured with traditional CMs, or existing table-based abstractions, which are designed to match only certain regions of the device characteristics. This includes the following.
    - a) Properly modeling for device geometrical effects (e.g., nonrectangular device shapes).
    - b) Capturing complex interaction of different process parameters whose effects may not be well modeled in existing device models like stress, mobility mismatch, and other device characteristics.
  - 3) With the increase in variability, process variations often stretch beyond the modeling limits, and traditional CM-based approaches often rely on extrapolations to account for such large variations. Whereas a Tfm

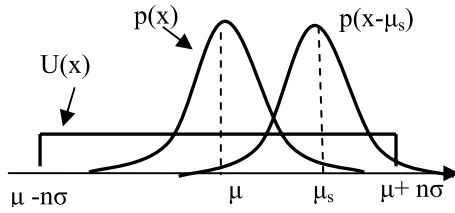


Fig. 4.  $p(x)$  is the natural distribution.  $U(x)$  is a uniform probability density function.  $p(x - \mu_s)$  is  $p(x)$  shifted to new center  $\mu_s$  for importance sampling [2].

approach captures the true and physical device characteristics. This is critical for rare event estimation.

- 4) TfM approach can enable on the spot process/memory design yield optimizations even if real hardware is not available. For example, we can study the stability of the SRAM cell if we decide to change some halo implants, special types of doping, and hence forth.

### A. Fast Statistical Sampling Method

Fast Monte Carlo methods overcome the slow convergence problem compared with traditional Monte Carlo in relation to low fail probability estimation. These methods are therefore a preferred alternative when it comes to the yield analysis of large and repetitive array structures.

Other alternatives like model or sensitivity-based approaches [6] may also be considered but true statistical sampling is more accurate as it evaluates the system at the sample points and provides true representation for the failure region. Model-to-hardware corroboration shows an excellent matching between importance-sampling-based methods yield estimation [1] and the true hardware yield. We therefore adopt the methodology in [1] as the core statistical engine for our TfM methodology. The methodology requires two sampling stages and is best shown in Fig. 4 below.

Stage 1: Uniform sampling is employed to identify most probable failure region and hence the center of gravity of fails of a given design metric.

Stage 2: Statistical sampling is performed around the center of gravity using shifted distributions. The shift is the center of gravity derived in step 1. The sample probabilities are unbiased to estimate the fail probability.

This is based on the fact that probability estimated with a natural distribution  $p(x)$  is equal to the probability of samples obtained from a biased distribution  $g(x)$  with adjusted weights

$$E_{p(x)}[\theta] = E_{g(x)}\left[\theta \cdot \frac{p(x)}{g(x)}\right]. \quad (1)$$

By focusing on critical fail regions, the number of sample points for estimating the fail probability  $P_f$  is reduced dramatically. For standard Monte Carlo methods, the required number of sample points,  $N$  is inversely proportional to the probability of fail  $P_f$  of the estimate. Hence, for a memory design that fails at five standard deviations, several millions of sample points are needed to estimate the corresponding probability of fail ( $P_f = 3e - 7$ ) with confidence. Methods

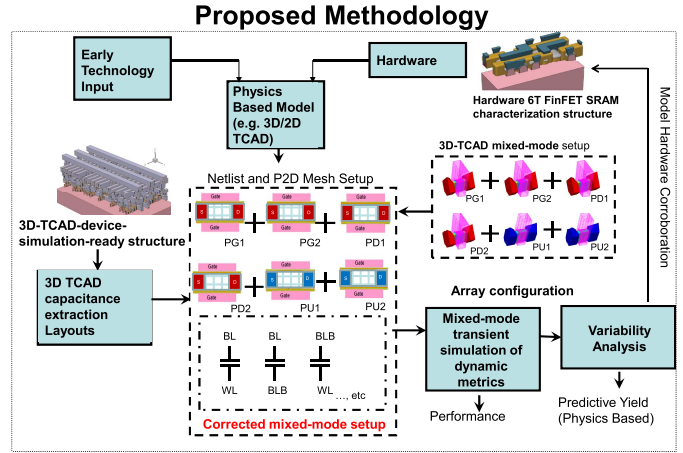


Fig. 5. Fast and accurate TfM flow-SRAM cell example.

like [2] require only a few thousand simulations to converge for the same estimate.

### B. Fast and Accurate Mixed-Mode Simulation Flow

Statistical simulations require accurate simulation engine. At the heart of the proposed statistical engine lies, the mixed-mode simulations engine [Fig. 3(b)] used to simulate the design dynamic margins. To bring the proposed flow to practicality, a primary goal is to optimize the mixed-mode simulation runtime while maintaining TCAD accuracy. Fig. 5 summarizes the proposed methodology flow.

The key ingredients of transient simulation are accurate device models and parasitics and our methodology relies on the following relevant features.

- 1) *Efficient TCAD Simulation*: A 2-D cross-sectional TCAD structure (P2-D) model generation flow is designed to enable fast and accurate mesh structures that capture the 3-D FinFET device characteristics and parasitics.
- 2) *3-D TCAD Layout-Aware Capacitance Extraction*: An advanced automated FinFET parasitic extraction method is adopted to extract the SRAM cell layout parasitics. The parasitics are in turn embedded in the netlist for purposes of more accurate mixed-mode dynamic margin simulations.

Hence, compensation capacitors,  $\Delta C_{\text{geom}}$  and  $\Delta C_{\text{ckt}}$ , are appended to the circuit (comprising of P2-D devices) nodes to account for internal device parasitics modeling mismatch as well as circuit/layout-dependent parasitics

$$C = C_{P2-D} + \Delta C_{\text{geom}} + \Delta C_{\text{ckt}} \quad (2)$$

where  $C$  is the total capacitance at a given node and  $C_{P2-D}$  is the device capacitance implicit to P2-D device model.  $\Delta C_{\text{geom}}$  represents the 3-D to P2-D device-geometrical capacitance difference (Section III), and  $\Delta C_{\text{ckt}}$  represents the layout parasitics (Section IV).

## III. EFFICIENT TCAD DEVICE SIMULATION

### A. P2-D Model

It is well known that 3-D simulations are computationally intensive especially when the simulations involve dynamic

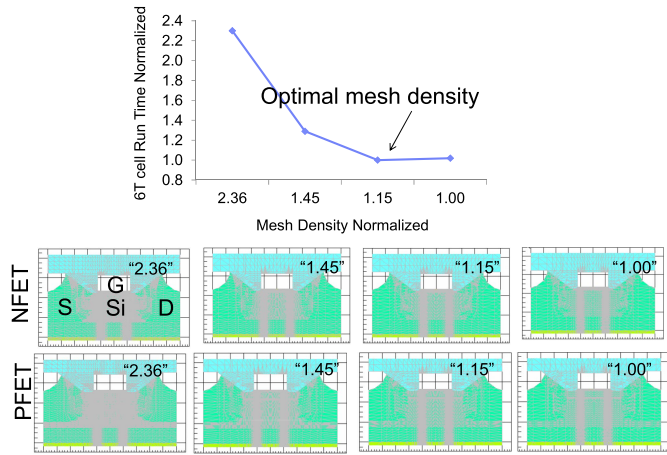


Fig. 6. Normalized 3-D dynamic margin simulation runtime. Decreasing mesh density (shown in gray) from left to right.

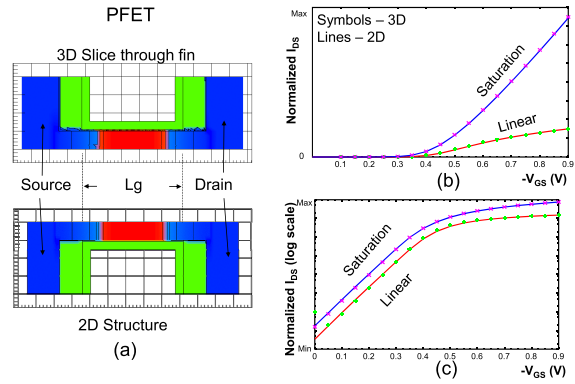


Fig. 8. (a) TCAD-simulated 3-D and P2-D structures. TCAD-predicted dc  $I$ - $V$  curves for pMOS FinFET match in superthreshold and subthreshold regions as illustrated in the normalized, (b) linear-scale, and (c) log-scale dc  $I$ - $V$  curves.

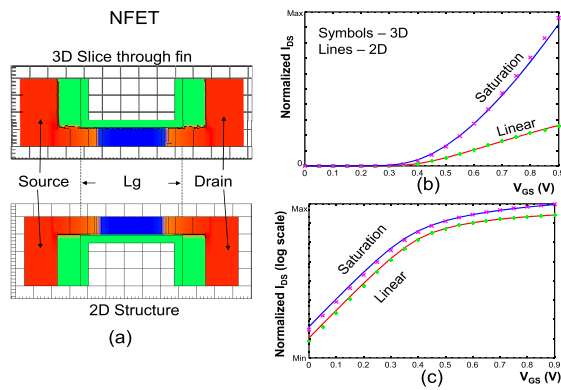


Fig. 7. (a) TCAD-simulated 3-D and P2-D structures. TCAD-predicted dc  $I$ - $V$  curves for nMOS FinFET match in superthreshold and subthreshold regions as illustrated in the normalized, (b) linear-scale, and (c) log-scale dc  $I$ - $V$  curves.

analysis. Mesh refinement techniques can help improve the runtime as shown in Fig. 6 but cannot eliminate most of the overhead.

To improve the simulation efficiency while maintaining accurate results, we propose to map the full 3-D device geometry physically to a 2-D cross-sectional TCAD structure (e.g., P2-D) by employing all device characteristics of the full device. The P2-D device structure is developed to obtain device output that is representative of the full 3-D device structure. Fig. 7(a) compares a slice of the full 3-D FinFET structure through the fin center and the P2-D structure. The structures and the doping profiles are similar. Physics models in 3-D and 2-D simulations include drift-diffusion, Fermi-Dirac statistics, quantum correction, generalized mobility model, Shockley-Read-Hall, Auger, surface recombination/generation models, and band-to-band tunneling. The 2-D lumped source and drain (S/D) resistance is scaled from 3-D by  $W_{\text{eff}}$ . Thus, as shown in Figs. 7 and 8 for the dc comparison, device characteristics are the same, and 2-D-based analysis is truly representative. 3-D and P2-D simulations were performed using Fielday [7].

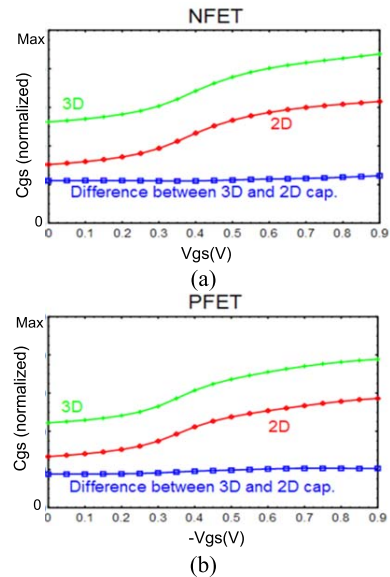


Fig. 9. Fielday-simulated  $C_{gs}$  versus  $V_{gs}$  in the (a) n- and (b) p-type FinFETs.

### B. P2-D Model—Compensation Caps

Due to the structural difference in 3-D and 2-D, the discrepancy of parasitic capacitance in 3-D and 2-D must be modeled to achieve equal ac performance. Fig. 9 shows Fielday-predicted [7] gate-to-source capacitance ( $C_{gs}$ ) versus  $V_{gs}$  at  $V_d = V_s = V_{\text{sub}} = 0$ . The  $C_{gs}$  difference between 2-D and 3-D structures is also depicted in the same figure. Interestingly,  $C_{gs}$  difference is independent of gate bias. We use the same mesh and dopant profile in P2-D device. Thus, P2-D structure has the same bias dependence with 3-D device. In such a case, we can add lump capacitances in the gate-to-source and gate-to-drain for P2-D structure to compensate for the geometrical parasitic penalty in 3-D devices [8] and hence provide the same ac and transient simulations. From the figure, we can use  $C_{gs}$  and  $C_{gd}$  adds as 11.15 aF in NFET and 9.79 aF in PFET for P2-D devices. By simply considering this capacitance difference, Fig. 10(a) show a P2-D device with extrinsic geometrical

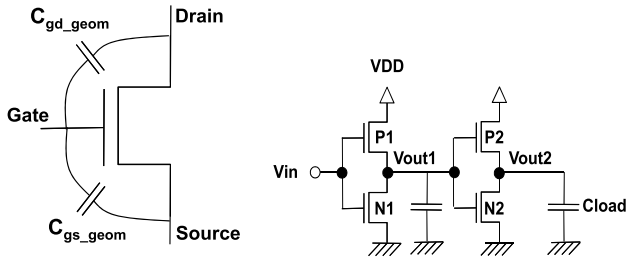


Fig. 10. (a) P2-D geometric compensation caps. (b) Two-stage inverter.

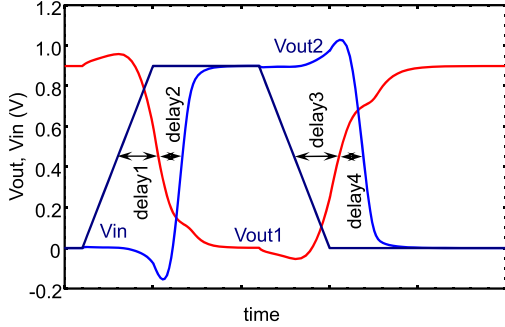


Fig. 11. Illustration of the delays to be measured.

compensation capacitors gate-to-source ( $C_{gs\_geom}$ ) and gate-to-drain ( $C_{gd\_geom}$ ) representing the front end device capacitance matching.

Using the resultant capacitance compensated devices, we demonstrate the accuracy of the proposed P2-D model by evaluating the delays of the simple two-stage fanout-of-1 inverter chain in Fig. 10(b). Aggressive rise/fall-times are used to increase the sensitivity of the delay to the device parasitics. Fig. 11 shows a set of four delays.

- 1) Delay 1: V<sub>out1</sub> falling edge.
- 2) Delay 2: V<sub>out2</sub> rising edge.
- 3) Delay 3: V<sub>out1</sub> rising edge.
- 4) Delay 4: V<sub>out2</sub> falling edge.

Simulations indicate an excellent matching (<2% error) for the tuned P2-D-based inverter chain delays compared with the 3-D-based inverter chain. Table I presents the normalized delay. Fig. 12 shows perfect waveform matching of the output node simulations of the compensated P2-D model to the 3-D model simulations. The figure also highlights the response mismatch in the absence of compensation caps and hence their criticality. Fig. 13 shows the equivalent P2-D-based schematics of an SRAM cell including the compensation caps. Hence, it represents P2-D structure-based SRAM cell with calibrated compensation capacitance component (FE device cap matching); e.g.,  $\Delta C_{gs\_geom\_PR}$  is the source-to-gate tuning capacitor adder for the right pull-up device PR.

Table II summarizes the average error for both the inverter chain and SRAM cell static and dynamic responses. By static simulation, we are referring to the steady-state simulations with wordline initially zero then turned on for long time to capture steady-state node upset. The dynamic response represents a read after write simulation. The recorded error on average is less than 3%. Most importantly, the resultant configuration

TABLE I  
NORMALIZED DELAY (TO THE MINIMUM RECORDED 3-D DELAY).  
NOMINAL DEVICES WITH NO VARIABILITY

(delay normalized)	Delay1	Delay2	Delay3	Delay4
2D, no C <sub>gs</sub> , C <sub>gd</sub> adders	1.10	0.64	1.20	0.66
2D, with C <sub>gs</sub> , C <sub>gd</sub> adders	1.46	1.02	1.65	1.06
3D	1.41	1.00	1.62	1.04

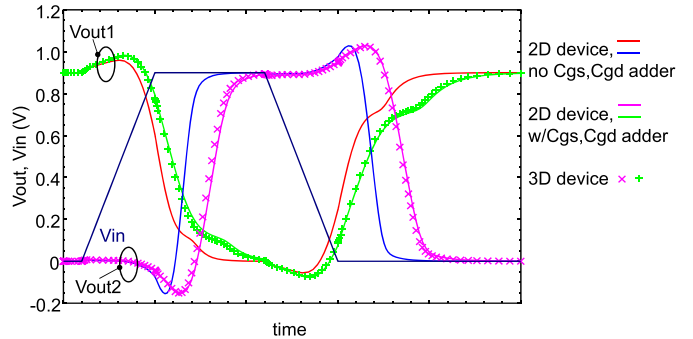


Fig. 12. Comparison of 3-D and 2-D with and without compensation caps added. Nominal simulation. Time scale in picoseconds.

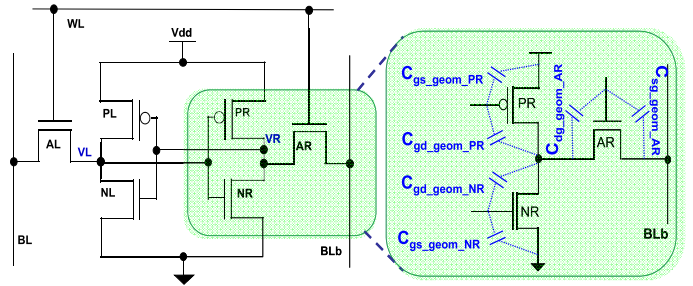


Fig. 13. Schematics of P2-D-based SRAM structures.

TABLE II  
3-D VERSUS P2-D SIMULATIONS. ACCURACY AND SPEED. STATIC SIMULATIONS MEASURE THE NODE UPSET (NOISE DURING READ) ERROR. DYNAMIC SIMULATIONS MEASURE 50% DELAY ERROR

	Full 3D TCAD (normalized runtime)	Efficient P2D (normalized runtime)	Average Error (P2D vs Full 3D)
SRAM Static Simulation	244.60	1	2.5%
SRAM dynamic Simulation (e.g., write performance)	276.88	1	3%
Logic Gate (inverter Chain)	206.01	1	2%

enables dynamic margin simulations with matched accuracy and 200× speedup compared with full 3-D simulations.

### C. Device Variability Modeling

For variability analysis, we leverage physical aspects of the FinFET structure such as the VGS dependence on surface

TABLE III  
WF SHIFTS APPLIED TO DEVICES IN FIG. 7

Device	Workfunction Shifts Applied
N1 (Stage 1 pulldown device)	+50mV
P1 (Stage 1 pullup device)	-50mV
N2 (Stage 2 pulldown device)	-50mV
P2 (Stage 2 pullup device)	+50mV

TABLE IV

NORMALIZED DELAY (TO THE MINIMUM RECORDED 3-D DELAY) FOR THE CIRCUIT OF FIG. 7. +/-50-mV WF SHIFTS APPLIED IN THE WORST CASE DIRECTIONS ACCORDING TO TABLE III

(delay -normalized)	Delay1	Delay2	Delay3	Delay4
<b>2D, with Cgs, Cgd adders</b>	1.84	1.02	2.07	1.02
<b>3D</b>	1.78	1.00	2.04	1.01

potentials, front/back-gate voltages, as shown in (3) [9]

$$V_{GS} = \psi_{sf} + \frac{1}{1+r} \left[ (V_{FBf} + rV_{FBb}) - \left( \frac{Q_{cf}}{C_{of}} + r \frac{Q_{cb}}{C_{ob}} \right) - \left( \frac{Q_b}{2C_{of}} + r \frac{Q_b}{2C_{ob}} \right) \right] \quad (3)$$

where  $r$  is a gate-gate coupling factor expressed as

$$r = \frac{C_b C_{ob} Q_{cf}}{C_{of} (C_b + C_{ob})} \cong \frac{3t_{oxf}}{3t_{oxb} + t_{Si}} \quad (4)$$

$V_{GS}$  and  $V_{GBs}$  are the gate-to-source voltages,  $V_{FBf}$  and  $V_{FBb}$  are the front- and back-gate flat-band voltages,  $\Psi_{sf}$  and  $\Psi_{sb}$  are the front- and back-surface potentials,  $Q_{cf}$  and  $Q_{cb}$  are the front- and back-surface inversion charge densities,  $Q_b = -qN_A t_{Si}$  is the depletion charge density,  $C_{of} = \epsilon_{ox}/t_{oxf}$  and  $C_{ob} = \epsilon_{ox}/t_{oxb}$  are the front- and back-gate oxide capacitances, and  $C_b = \epsilon_{Si}/t_{Si}$  is the depletion capacitance. In our FinFET structure,  $V_{FBf}$  and  $V_{FBb}$  are the same, thus  $\Delta V_{FB} = \Delta V_{GS}$ . Note that,  $\Delta V_{FB}$  can be the variation of gate WF where the fixed gate-insulator charge is negligible. For the symmetrical gate structures,  $V_{FBf} = V_{FBb}$  in (3) and  $r = 1$  in (4) [9], thus, the variation of WF is  $\Delta V_{GS}$  in (3). Hence,  $\Delta WF$  can capture the device threshold-voltage fluctuation ( $\Delta Vt = \Delta V_{GS}$ ). Usage of  $\Delta WF$  to model device variability eliminates the complexity that arises in traditional approaches similar to Fig. 1 due to atomistic simulations and device mesh regeneration.

Finally, we test the accuracy of the compensation cap calibration of the P2-D-based design in the presence of variability. We repeat the simulations for the circuit of Fig. 13 in the presence of +/-50 mV WF shifts applied in the worst case directions according to Table III. We present the results in Table IV and record matching waveforms with less than 3% relative error.

#### IV. AUTOMATED LAYOUT CAPACITANCE EXTRACTION

To determine dynamic margins of the FinFET SRAM array, it is essential to capture bitcell parasitic capacitances

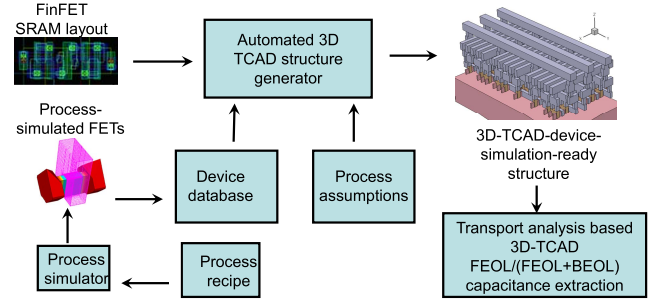


Fig. 14. Capacitance extraction flow for FinFET SRAM.

accurately. The adopted parasitic extraction flow offers a unique opportunity to enable enhanced accuracy by modeling back-end-of-the-line (BEOL) effects and front-end-of-the-line (FEOL) interaction effects. To capture layout-specific contributions, the methodology relies on a modification of the FinFET capacitance extraction technique tested for the first time here on industrial designs and processes, and verified through hardware at IBM for 32 nm [10]. Automated structure generation is used for nonplanar devices. The contributions of different layers to bitline and wordline caps are estimated taking into consideration the impact of fin spacing or fin pitch. Storage node capacitors are also revisited. The derived capacitances are then embedded in the mixed-mode schematic simulation flow that is used to predict the dynamic behavior for different physical design considerations including low voltage operation. Fig. 14 shows the proposed extraction methodology.

#### A. 3-D Layout-Specific Extraction for FinFET SRAM

FEOL/(FEOL + BEOL) extraction is a major problem for nonplanar devices [11]. Special rules are needed to recognize FinFETs. Specifically there is a need to identify real versus dummy fins and construct a variety of epitaxial-silicon shapes grown on S/D regions. Here is a summary of underlying features/steps applied.

- 1) We first generate the device structures and then merge them with the other structures where S/D are shared. This involves completing the layouts incrementally, then adding the epilayers appropriately.
- 2) Since segregated FEOL/BEOL capacitance extraction approaches are questionable for FinFETs, we perform 3-D-TCAD capacitance extraction on  $3 \times 2$  (FEOL + BEOL) FinFET SRAM array structures (Fig. 13). Using a nominal sub-22-nm process recipe in Sentaurus Process [12], a FinFET database consisting of n/p FinFETs is built. Individual FinFETs are automatically recognized from the input SRAM layout as the intersection of fin, gate conductor, and active regions, by an automated 3-D-TCAD structure generator, which synthesizes the corresponding device-simulation-ready structure with the aid of FEOL/BEOL process assumptions.
- 3) We then perform transport-analysis-based extraction on the structure to obtain the parasitic capacitances. We refine the device meshes accordingly, as shown in Fig. 15. To speedup extraction, metal meshes are constrained to the equipotential surfaces of the metal layers.

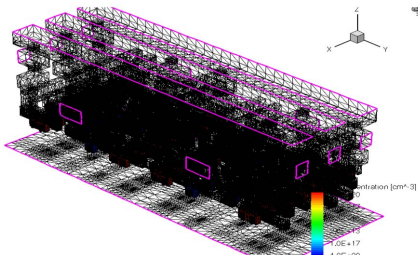
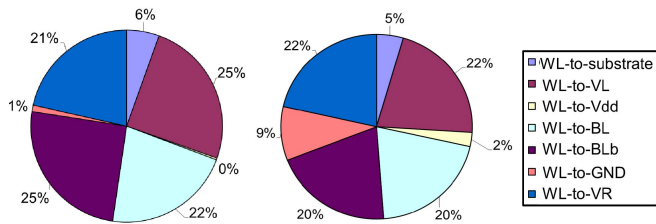
Fig. 15. Refined mesh for  $3 \times 2$  cell structure.

Fig. 16. Contributions to wordline capacitance (cell in Fig. 13).

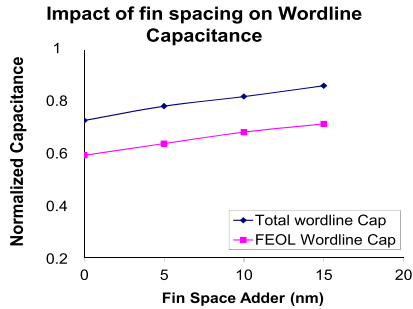


Fig. 17. Impact of fin space adder on wordline capacitance.

- 4) To separate the BEOL from FEOL structures, two setups are built. The first includes front-end layers up to the tungsten contacts layer, and the second includes all the metal layers. The first provides the FEOL parasitics and interactions, and the latter provides (FEOL + BEOL). The difference results in the layout-specific (BEOL) parasitics that are key components of wordline and bitline capacitance. Finally, FEOL interactions are uniquely computed for all specified nodes.

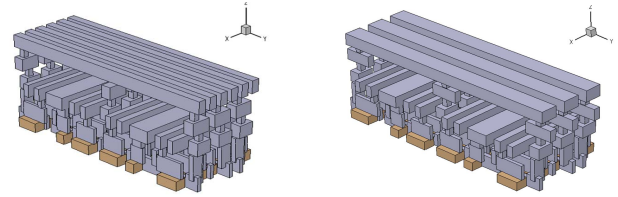
An example breakdown of the contributions of the layout parasitics to bitline and wordline capacitance, is shown in Fig. 16 for wordline case, and can guide design layout optimizations of the cell.

We study in Fig. 17 the impact of the fin pitches,  $p_{fin} = p_{fin0} + \Delta p_{fin}$ , on bitline and wordline capacitance by moving all FEOL and BEOL shapes with the fin space adder ( $\Delta p_{fin}$ ).

Fig. 18 shows the two implementations of the same design where a 3.4%(1.5%) reduction in the bitline (wordline) capacitance is demonstrated. The  $3 \times 2$  cross section with six metal lines has larger BEOL effects. However, it has smaller total capacitance due to an optimized front-end.

## V. ANALYSIS

In this section, we study FinFET SRAM dynamic margin yield [13] using the proposed physical-based

Fig. 18. Small and large bitline capacitance implementation of the same cell. Cross section is for  $3 \times 2$  cells.

mixed-mode design flow. We demonstrate the accuracy of the P2-D model and perform static noise margin (SNM) [13] model-to-hardware corroboration. Here, we first highlight the discrepancy between CMs and the physics-based approach.

### A. SPICE CMs Versus 3-D TCAD Devices

Experiments show that CM current-voltage ( $I-V$ ) versus  $V_{gs}$  and  $V_{ds}$  device characteristics match well at nominal conditions with respect to 3-D TCAD simulations. However, the  $C-V$  characteristics of SPICE CM versus 3-D TCAD model deviate under certain operating conditions. We study those characteristics for total device gate and drain capacitance ( $C_{gg}$ ,  $C_{dd}$ ) as function of gate and drain voltages, respectively (Fig. 19 for  $C_{gg}$ ,  $C_{dd}$  results were similar). For nominal device, we notice deviation in the CM near low voltage operation. We then vary the threshold voltage by varying the WF of TCAD device in the mixed-mode simulation. We notice that this trend extends to higher supply voltage regions as the threshold voltage variation for the device increases from three to six standard deviations. The deviation in capacitor values can be more than 8% in certain regions. Overall, the lack of accurate parasitics modeling makes it difficult for the traditional CM simulation approach to predict proper functionality and enable design optimizations. Note that, the differences highlighted here are mainly due to device intrinsic parasitics only (no layout effects).

### B. SRAM Dynamic Margins: Simulation Setup and Yield

Dynamic margins of the design rely on the transient behavior of the circuit [11]. Three factors affect the delay and noise margins of the cell:  $I-V$ ,  $C_{gg}$ , and  $C_{dd}$ . For SRAM cells, the drain capacitance is particularly important since the storage node shares the drains of three devices. In the writability as well as read stability simulations (dynamic noise margins),  $C_{dd}$  is very important;  $C_{dd} = C_{dx} + C_{db} + C_{dg}$  is dominated by  $C_{dg}$ . Fig. 20 shows the left storage node (VL) capacitance. It also illustrates circuit waveforms during write operation. Our goal is to properly match the parasitics according to (1) to capture 3-D internal parasitics as well as circuit and layout interaction effects.

Fig. 21 shows the discrepancy in dynamic margins between SPICE compact models and TCAD simulation flow. The mismatch near low voltage operation exceeds 20% for nominal devices. With added variability, the error can get larger. The analysis is for write 0 on node VR, and write 1 on node VL. Fig. 22 shows the 2-D/3-D TCAD matching waveforms for  $V_{dd} = 0.9$  V.

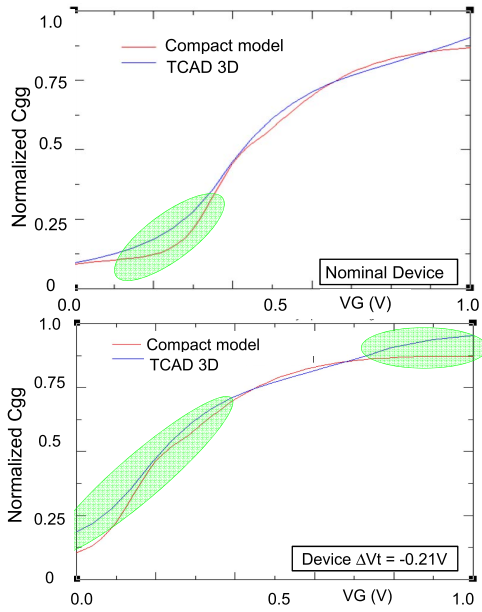


Fig. 19. CV comparison for the gate capacitance between TCAD and CM for nominal and mismatched devices.

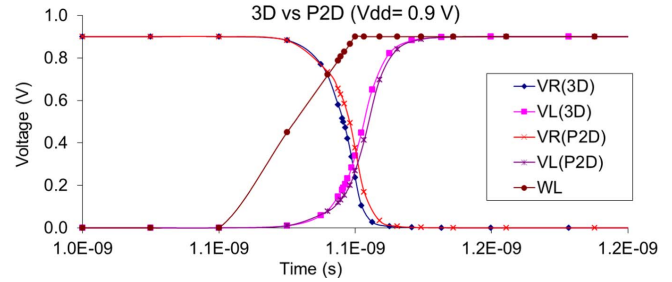


Fig. 22. Performance comparison of 3-D versus P2-D TCAD shows perfect match.

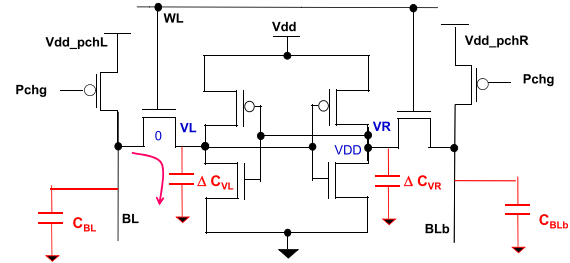


Fig. 23. Mixed-mode simulation flow schematic for modeling dynamic read noise margins.

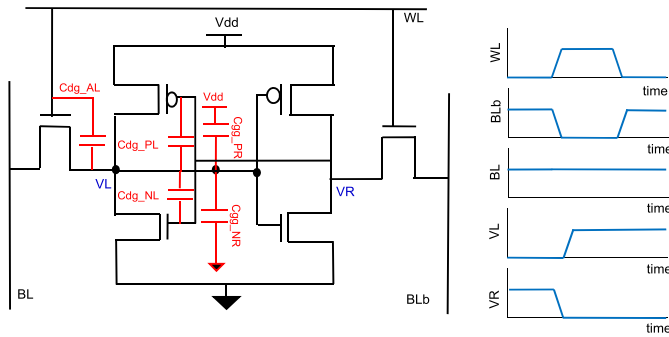


Fig. 20. SRAM write operation. The parasitic capacitance of a storage node (e.g., VL) is highly dependent on  $C_{dg}$  and  $C_{gg}$ , which in turn are function of compensation capacitors ( $C_{gd}$  and  $C_{gs}$ ).

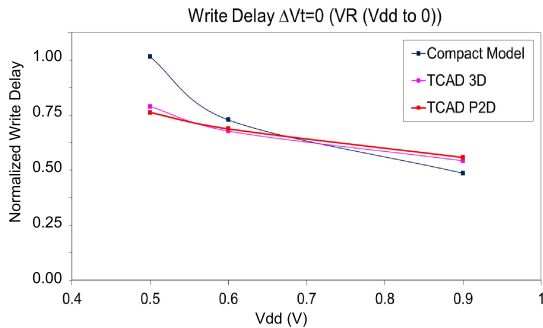


Fig. 21. Comparison of TCAD versus SPICE CM.

Eventually, lumped capacitors can be added to critical nodes to model proper parasitic effects. Fig. 23 shows the setup for dynamic read noise margin, stability, and analysis; particularly, precharge devices are necessary for floating bitline analysis. Layout effects (taking into consideration number of cells/bitline, cells/wordline) are modeled in terms of lumped wordline, bitline, and internal node parasitics. The setup helps emulate floating bitlines conditions for the mixed-mode simulations.

The proposed TfM P2-D circuits are then used to study the dynamic read and write margins for different bitline loads and cell designs. Figs. 24 and 25 show the TfM yield for dynamic stability and writability yields of state-of-the-art dense sub-nm SRAM cell. The difference is quite obvious when the TfM estimates are compared with traditional CM-based yield estimates. Consistently, the CM-based approach brings in unwanted pessimism. Note that, there are many nonlinear mismatches that come into play when dealing with variability due to the semiempirical nature of SPICE simulations. For TfM, the physical parameter variation has different implications on the device strength as opposed to simple  $Vt$  adder. Modifying WF will automatically amplify the variation in other dependencies. Whereas in SPICE, the other dependencies will not be amplified because the equations are semiempirical, add to this the fact when variability goes out of range, the error will be amplified. Finally, the runtime improvement is significant, and a single yield estimate can be run within a few hours to estimate six sigma with confidence for the cell dynamic margin using TfM approach.

Finally, a key feature of the TFM methodology is the ability to predict yield trends for new processes whose CMs have not been developed. This is critical for technology/circuit co-design optimization. Fig. 26 compares the stability and writability yields of an SRAM design for two different process corners. The process recipe/doping is used to modify the WF in the direction of improved yield. P2-D predicts the improvement accordingly. It is obvious that process 2 is favorable.

### C. P2-D SNM Hardware Corroboration

SNM hardware measurements of a  $0.06\text{-}\mu\text{m}^2$  FinFET SRAM were obtained and the P2-D TCAD model is perfectly matched with the measured SNM curves. The resultant model was effective at predicting all the hardware-based SNM

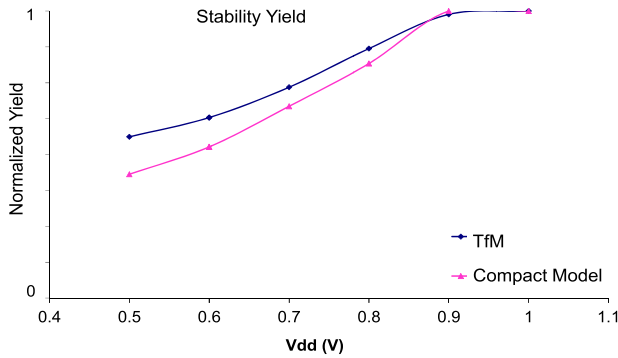


Fig. 24. Tfm (P2-D) versus CM-based yield—stability.

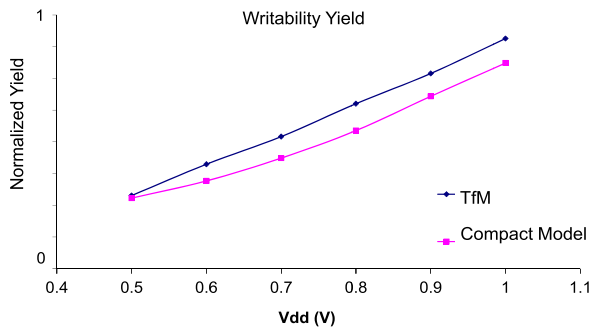


Fig. 25. Tfm (P2-D) versus CM-based yield-writability. Results correlate with higher % of error at 0.9 V.

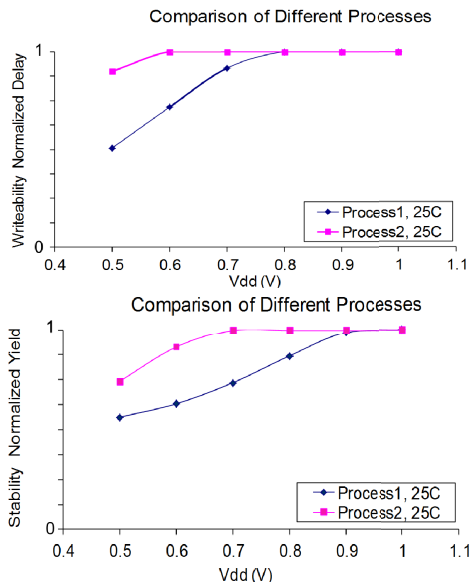


Fig. 26. Writability and stability comparison for two process corners. Simulations possible only via Tfm prior to CM generation.

measurements over a wide range of process corners, voltage, and temperature operating conditions including low voltage operation down to 0.4 V Vdd. Fig. 27 shows the case for  $V_{dd} = 0.9$  V as an example. The physics-based TCAD simulations truly rendered more accurate and predictable. Hence, the method can be effectively applied to early prehardware stages as well as updated-process/device in consecutive design cycles.

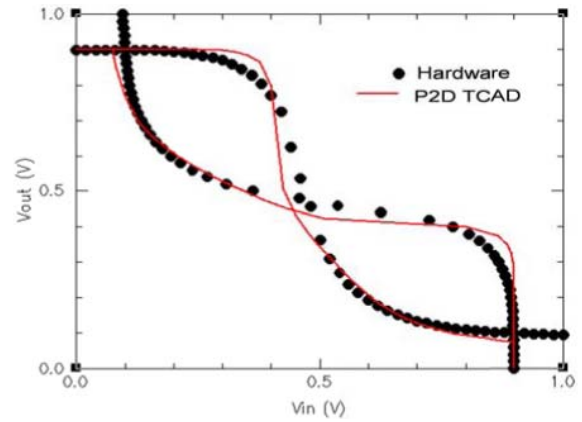


Fig. 27. Butterfly curves of  $0.06\text{-}\mu\text{m}^2$  cell. SNM curves match graph for the calibrated P2-D devices.

## VI. CONCLUSION

We presented a pragmatic physics-based mixed-mode statistical simulation methodology for the first time. The runtime, which was impractical for statistical dynamic margin analysis before is made feasible via the newly developed P2-D meshes with enhanced parasitic modeling. The P2-D model speeds up dynamic margin simulations, and brings Tfm to practicality reducing the simulation runtime from several months to hours for an eight-transistor design. At the circuit level, cell capacitances, which are a must for accurate circuit and layout interaction representation, are generated using automated structure generation. We evaluated the methodology using FinFET process technology as a vehicle and observed orders of magnitude speed improvement for 3-D FinFET-based SRAM cell design. In addition, we showed that SPICE CMs, which are derived from the 3-D FinFET models, deviate under statistical variation conditions.

## REFERENCES

- [1] A. Singhee and R. Rutenbar, "Statistical blockade: A novel method for very fast Monte Carlo simulation of rare circuit events, and its application," in *Proc. IEEE Des. Autom. Test Conf.*, Mar. 2007, pp. 1–6.
- [2] R. Kanj, R. Joshi, and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare fail events," in *Proc. IEEE Des. Autom. Conf.*, Jun. 2006, pp. 69–72.
- [3] C. Dong and X. Li, "Efficient SRAM failure rate prediction via Gibbs sampling," in *Proc. IEEE Des. Autom. Conf.*, Jun. 2011, pp. 200–205.
- [4] D. E. Hocevar, M. R. Lightner, and T. N. Trick, "A study of variance reduction techniques for estimating circuit yields," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 2, no. 3, pp. 180–192, Jul. 1983.
- [5] Synopsys Inc., Mountain View, CA, USA. (2013). *MEDICI: 2-D Device Simulation*. [Online] Available: <http://www.synopsys.com>
- [6] C. Wann *et al.*, "SRAM cell design for stability methodology," in *Proc. Int. Symp. VLSI Technol., Syst. Appl.*, Aug. 2005, pp. 21–22.
- [7] E. M. Buturla, P. E. Cottrell, B. M. Grossman, and K. A. Salsburg, "Finite-element analysis of semiconductor devices: The FIELDAY program," *IBM J. Res. Develop.*, vol. 44, no. 22, pp. 142–146, 2000.
- [8] W. Wu and M. Chan, "Analysis of geometry-dependent parasitics in multifin double-gate FinFETs," *IEEE Trans. Electron Devices*, vol. 54, no. 4, pp. 692–698, Apr. 2007.
- [9] K. Kim and J. Fossum, "Double-gate CMOS: Symmetrical- versus asymmetrical-gate devices," *IEEE Trans. Electron Devices*, vol. 48, no. 2, pp. 294–299, Feb. 2001.
- [10] A. Bhoj *et al.*, "Hardware-assisted 3D TCAD for predictive capacitance extraction in 32nm SOI SRAMs," in *Proc. IEEE Int. Electron Device Meeting*, Dec. 2011, pp. 34.7.1–34.7.4.

- [11] E. Karl *et al.*, "A 4.6GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active VMIN-enhancing assist circuitry," in *Proc. IEEE Int. Solid State Circuits Conf.*, Feb. 2012, pp. 230–232.
- [12] (2014). *An Advanced 1D, 2D, and 3D Process Simulator* [Online]. Available: <http://www.synopsys.com/TOOLS/TCAD/.../Pages/SentaurusProcess.aspx>
- [13] R. Joshi, R. Kanj, S. Nassif, D. Plass, Y. Chan, and C.-T. Chuang, "Statistical exploration of the dual supply voltage space of a 65nm PD/SOI CMOS SRAM cell," in *Proc. IEEE Eur. Solid State Device Rel. Conf.*, Sep. 2006, pp. 315–318.
- [14] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid State Circuits*, vol. 22, no. 5, pp. 748–754, Oct. 1987.

**Rajiv V. Joshi** (F'01) is a research staff member at T. J. Watson research center, IBM. He received his B.Tech I.I.T (Bombay, India), M.S (M.I.T) and Dr. Sc.(Columbia University). His novel interconnects processes and structures for aluminum, tungsten and copper technologies which are widely used in IBM for various technologies from sub-0.5 $\mu$ m to 14nm. He received 3 Outstanding Technical Achievement (OTAs), 3 Corporate Patent Portfolio awards for licensing contributions, holds 54 invention plateaus and has over 195 US patents and over 350 including international patents. He has authored and co-authored over 165 papers. He is a recipient of 2013 IEEE CAS Industrial Pioneer award and 2013 Mehboob Khan Award from Semiconductor Research corporation. He is Distinguished Lecturer for IEEE CAS and EDS society. He is IEEE and ISQED fellow and distinguished alumnus of IIT Bombay.

**Keunwoo Kim** (S'98–M'01–SM'06) received the B.S.(Sungkyunkwan University, Seoul, Korea, 1993) and the M.S. and Ph.D. degrees in EE (University of Florida, Gainesville, in 1998 and 2001). He was a Research Staff Member at IBM T. J. Watson Research Center, from 2001–2013. He worked on the design of high-performance and low-power microprocessors, novel VLSI circuit techniques, scaled and exploratory CMOS performance/power evaluation, and physics/modeling for bulk-Si, SOI, strained-Si, SiGe, hybrid orientation/device, and double-gate technologies. He received five invention achievement awards from IBM. In 2013, he joined Samsung Display Co. Ltd, Korea. He has over 85 papers in journals and conference proceedings. He holds over 25 U.S. patents, with several patents pending.

**Rouwaida Kanj** (M'99–SM'10) received the B.Eng. with high distinction from the American University of Beirut in 1998, and the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Illinois Urbana-Champaign in 2000 and 2004 respectively. She is currently an Assistant Professor at the American University of Beirut. Previously, she was a research staff member at IBM Austin Research Labs from 2004–2012. Dr. Kanj was a recipient of three IBM Ph.D. Fellowships, is the author of more than 40 technical papers, 20 issued patents and several pending patents. She also received the IEEE ISQED and IEEE ICCAD best paper awards, and is a senior member of IEEE.

**Ajay N. Bhoj** received his B.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Madras, India in 2007, and Ph.D. in Electrical Engineering from Princeton University, Princeton, NJ in 2013. He is currently with Intel Corporation, Hillsboro, OR. His interests span problems in device modeling, low power digital circuit design, image processing, computer vision, and computational lithography.

**Matthew M. Ziegler** received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville in 2004. He is with IBM T. J. Watson Research Center, NY since 2004. His research has focused on VLSI design productivity and low power design. He has developed design methodologies as well as power reduction and circuit optimization techniques used throughout IBM. He has directly contributed to the design of POWER, z, and BlueGene processor families. Dr. Ziegler has published over 30 peer reviewed papers in the areas of low power design, design productivity, arithmetic circuits, and nanotechnology.

**Phil Oldiges** (M'99–SM'12) received the BS in Physics from Thomas More College in 1981, and the M.S. and Ph.D. degrees in Electrical Engineering from Cornell University in 1984 and 1988, respectively. From 1984 to 1986, he was a visiting research scientist at Toshiba Corporation in Kawasaki, Japan. From 1988 to 1993, he worked at Sony Corporation, Atsugi, Japan. From 1993 to 1998, he worked in the TCAD group at Digital Equipment Corporation, MA. Currently, he is manager of the research TCAD group with the IBM Systems and Technology Group in Hopewell Junction, NY, and is responsible for front end process and device models for the 7nm technology node and beyond. He is a senior member of the IEEE, New York Academy of Sciences, tau beta pi, and sigma pi sigma.

**Pranita Kerber** received her M.S. and Ph.D. degrees in Materials Science and Engineering from Carnegie Mellon University, PA in 2006 and 2008, respectively. In 2008, she joined IBM Corp. as a Research Staff Member where she conducts research in process and device simulation of advanced semiconductor devices and circuits. She has received more than 40 IBM Invention Awards, has co-authored more than 40 technical peer reviewed papers, and is a reviewer for numerous IEEE publications.

**Robert Wong** joined IBM Microelectronics Division in 1966. He received his Master's degree in Solid State Physics from Syracuse University, and completed the Ph.D. program in Biophysics in 1970. He had worked on Functional Memory, 4 bit microprocessor, and chip set designs in Bipolar and CMOS technology for IBM systems. He taught electronics in Chinese University of Hong Kong during the IBM sabbatical leave in 1981–82. He was design leader of the last bipolar SRAM chip in 'record' density of 128K in 1992. He then joined the Flash Memory Group and led the design of the 48KB control store for small systems until 1996, when he started the journey of SRAM scaling from 130nm to the present 7nm. He had proposed the spec for SRAM operation margins in terms of expected failure rate since 130nm. That had been the main tool to diagnose the SOI floating body problems in the 90nm game chip for XBOX & play-station. SRAM operation margins have become standard metrics for product characterization and debug. He is one IBM master inventor with 59 patents.

**Terence Hook** is a graduate of Brown University and earned his Ph.D. in Electrical Engineering from Yale University. At IBM since 1980, he has worked on technology integration and device design for bipolar, BiCMOS, and CMOS technologies as well as process-induced charging, and most recently has focused on fully-depleted devices in both planar and FinFET varieties. He works closely with teams in East Fishkill and Albany, New York, as well as Essex Junction, Vermont. He holds some dozens of patents and has authored more than 80 technical papers.

**Sudesh Saroop** received his B.S. (City University of New York in 1989) and his M.S. and Ph.D. (1995 in Physics, and 1999 in EERensselaer Polytechnic Institute, Troy, NY). He joined IBM East Fishkill in 1989, where he was one of the first defect characterization engineers responsible for establishing and automating optical and SEM-based defect detection as a yield predictor in the Advanced Semiconductor Technology Center on 350 nm logic, DRAM, and SiGe products. He later worked on 200 nm, 140 nm, and 100 nm DRAM device design, simulation, and functional characterization, 90 nm logic FET modeling, 65nm and 32nm SOI SRAMs, and is currently leading 14nm SRAM development. He is currently with the IBM Semiconductor Research and Development Center, Hopewell Junction, NY.

**Carl Radens** received BA, Physics (Oberlin College, 1983) and Ph.D. (University of Cincinnati, 1990). He is a Senior Technical Staff Member (STSM) at the IBM Semiconductor Research and Development Center (SRDC) working in the area of technology development.

**Chun-Chen (Frank) Yeh** received the Ph.D. degree in Electrical Engineering from Yale University in 2008. Since then he has joined IBM Research at Albany Nanotech focusing on advanced CMOS technology R&D. He is the recipient of IBM's Eminence and Excellence award for the achievement in extending the silicon device scaling, and he has been elected as IBM Master Inventor for the outstanding contribution to corporate's IP portfolio. His current research interest is in the transistor architecture design for power/performance enhancement in 10nm technology regime and beyond.