Simulation of Semiconductor Processes and Devices 2016
Edited by E. Bär, J. Lorenz, and P. Pichler

(Invited) Future Perspectives of TCAD in the Industry

Terry Ma*, Victor Moroz*, Ricardo Borges*, Karim El Sayed*, Plamen Asenov†, and Asen Asenov†
*Synopsys Inc., Mountain View, California, USA
†Synopsys Inc., Glasgow, Scotland, UK
terry.ma@synopsys.com

***Abstract*—For the better part of the last 20 years, simulation of semiconductor processes and devices had been the main focus in modeling planar CMOS transistors. The introduction of FinFETs in 2010, along with the increasing use of non-Si materials added much complexity and cost in technology development. With multiple device architectures and material options to consider, TCAD has evolved from focusing primarily on process and device simulation of single devices to multiscale simulation of multiple devices including novel materials to analyze variability and technology impact on design. In this paper, we offer our perspectives of TCAD for its expanded role in pathfinding for early selection of technology choices and generation of pre-wafer Process Design Kit (PDK).**

Keywords—TCAD, simulation, pathfinding, DTCO, PPA, PDK, SPICE, library cells, 14nm, 5nm, FinFET, nanowire, MOL, BEOL, circuit, Shift Left

I. INTRODUCTION

Over the last two decades, Technology Computer Aided Design (TCAD) has proven its value and steadily expanded its role in technology development [1]. As CMOS scaling continues, semiconductor devices are being pushed to the physical limit [2], requiring advanced physics such as quantum mechanics to be included in modeling these advanced semiconductor transistors. Thus far, the TCAD community has responded and answered the challenges of providing tools and novel techniques to address the increasing complexity of modeling semiconductor transistors with new architectures, transport phenomena, switching mechanisms, and materials [3] [4].

Figure 1 shows an example of using advanced TCAD tools and modeling techniques to compare the normalized transistor strengths of a planar CMOS transistor, FinFET, and nanowire [5]. Beside simulating nominal performance of single transistors, TCAD has recently been pushed to handle the modeling of multiple transistors (e.g. SRAM) in a single simulation to include proximity effects on device variability and parasitics such as Middle-of-Line (MOL) capacitance in FinFETs. To take it one step further, trending now in the industry is the so-called "Design-Technology Co-Optimization (DTCO)" for early selection of technology options as process and design are no longer isolated, sequential development activities, leading to the requirement of simulating test circuits in pre-wafer stage to understand their interactions. While DTCO is a broad term, in the following sections, we will share our perspective with two TCAD-Based DTCO examples, showing how TCAD can address the challenge of "Shifting Left" of the development cycle (see Figure 2) for pathfinding and contributing to the early Process Design Kit (PDK) development.

II. 5NM POWER-PERFORMANCE-AREA ANALYSIS

With transistors scaled down to 5nm design rules, the best Power-Performance-Area (PPA) trade-off can only be achieved by a holistic analysis that goes beyond a single transistor and includes MOL and Back-End-Of-Line (BEOL) resistances and capacitances. In this section, we will demonstrate a TCAD-only DTCO approach of analyzing the power-performance-area tradeoff for a 5nm design.

In this flow, standard cell performance is evaluated by building a library cell structure and calibrating device models to quantum transport physics, and the entire cell is treated as a single simulation domain, instead of resorting to SPICE analysis because reliable SPICE models for novel devices may not yet be available at the early stages of technology development.

Figure 3 shows a 9-track tall 2-NAND logic cell with a gate pitch of 32nm, metal pitch of 24nm, and fin pitch of 18nm. Several versions of this library cell have been implemented with 45nm tall FinFETs and lateral nanowires. A 3D structure built by Process Explorer is depicted in Figure 4, with two fins and two levels of

978-1-5090-0818-6/16/$31.00 ©2016 IEEE

stacked lateral nanowires, where each transistor has four nanowires connected in parallel.

Transient switching behavior of the whole cell that includes several transistors and interconnects, as well as an external load of fan-out of 2 and 70 metal pitches long BEOL wire is analyzed in Sentaurus Device to obtain performance and power consumption for different technology options. Comparative analysis of five technology options is summarized on Figure 5, where power-performance trade-off curves for each transistor architecture are obtained by changing power supply voltage in the range from 0.75 V to 0.6 V.

At 5nm design rules, the PPA trade-off is dominated by MOL capacitances and shows that fin depopulation from 2 fins to 1 fin improves power consumption by 30%. Switching from FinFETs to nanowires provides further gains, about 50% for two lateral nanowires per transistor. Figure 6 illustrates transistor strength and pin capacitance for each of the five technology options.

In terms of first-order effects, the benefits of fin depopulation and transition from fins to nanowires can be explained by the dominant role of capacitance vs. transistor strength. Consider that switching delay is roughly proportional to CV/I, and that dynamic power consumption is proportional to CV^2. Therefore, transistor's driving current (I) only affects the switching speed, whereas the load capacitance affects both the switching speed and the power consumption.

This 5nm PPA analysis was performed for a given set of material properties. Analysis of the impact of different material properties on PPA trade-off opens a much wider optimization window and enables trading off between design rules, transistor architecture, and material processing in search of the best PPA for a particular chip design application.

III. 14NM TCAD-SPICE

In this section we present an example of a TCAD-Based early SPICE model extraction and PDK development for 14nm FinFET technology [6].

Global variations are modeled via different process splits accounting for the systematic variations in implant dose, geometrical critical dimensions depending on the location of the wafer as well as layout dependent effects [7]. The ranges and distribution of these process parameter splits are user-inputs that can be obtained either via extrapolation from previous technology nodes or obtained from equipment vendors. This information can be extremely technology, foundry, process and even fab specific. However, understanding this process space is critical to both technology development and yield ramping. Based on this information, a design of experiments (DoE) is defined which, in this case, consists of a set of 50 representative "device process splits." These 50 process splits consist of 25 n-MOS devices and the corresponding 25 p-MOS devices to reflect 25 "CMOS process splits". The simple DoE grid in this example is shown in Figure 7. For each of these 50 device splits a full 3D process simulation is performed with Sentaurus Process, and subsequently a full set of IVs is simulated with Sentaurus Device. These 50 sets of IV curves serve as references for subsequent analysis steps.

The execution of this DoE is controlled by the framework tool Sentaurus Workbench, which enables efficient use of the available computer cluster to execute either all 50 splits – or some subset – in parallel on different computers within the cluster. Further Sentaurus Process and Sentaurus Device support multi-threading, here, 8 CPUs were used for all simulations.

To account for local variability we deploy the variability engine Garand. In a first stage, for each of the 50 device process splits Garand is calibrated against the reference IVs from Sentaurus Device. This includes density gradient (DG) quantum corrections, inversion charge calibration and mobility model calibration. This step is fully automated leveraging the auto-calibration technology of Enigma.

Enigma also drives Garand to generate, for each of the 50 samples, a statistical ensemble of 200 atomistically different devices, physically modeling the combined effect of all major sources of local variation. These include: random discrete doping fluctuations (RDF), gate edge roughness (GER), fin edge roughness (FER) and metal gate granularity (MGG) variability [8]. An example of one FinFET subject to all local variability sources is illustrated in Figure 8. Garand simulates a full set of IVs for each of these 'atomistically different' device realizations and is numerically optimized to execute each of these 3D device simulations in a very fast and efficient way, delivering a 100% convergence rate.

To enable the variability aware DTCO flow, Enigma automatically manages the massive simulation data flow and job scheduling tasks using an internal database.

Once all the target I-V/C-V characteristics are generated using physical TCAD simulation, hierarchical compact models can be extracted for each of the 50 device splits. This is a two-stage process, involving:

1) The extraction of 'uniform' or 'base' SPICE models for each point in the DoE space. The approach can be extended to a 'response surface' SPICE model to allow for off-DoE grid model generation and circuit simulation [9]. The whole process can be easily automated using a

robust compact model extraction strategy implemented in the compact model extraction tool, Mystic.

2) At the second stage, the statistical set of I-V characteristics at each node of the DoE are used to extract local 'statistical' models using a carefully selected subset of the compact model parameters. The procedure is outlined in [6], and the results of the extraction for one device are shown in Figure 9 comparing the distribution of key figures of merit obtained from the physical TCAD variability simulation and the extracted statistical compact model in nominal conditions.

These extracted SPICE model cards are then used to build a model card library. Based on this, the statistical circuit simulation tool RandomSPICE automatically populates the generic simulated circuit with unlimited statistical transistor models using the ModelGen technology. ModelGen handles correlated - non-Gaussian distributions, as well as the correlation between global or the local variability and n- and p-channel transistors.

For a given circuit, such as a library cell, RandomSPICE generates a (random) set of process parameters, based on the real parameter distributions as supplied through process simulation or measurements, and – using a response surface model approach – will generate the proper global SPICE model parameters to be used for all transistors in the given instance of the standard cell. RandomSPICE then accounts for the local variability by generating randomized versions of the SPICE model cards (as shown in Figure 9), in accordance with (interpolated) local parameter variations.

Subsequently, RandomSPICE calls HSPICE to perform a circuit simulation for each of these randomized circuit instances, stores all results in the database, and generates the relevant output statistics. In this example, the results of a 7-stage ring oscillator (RO), which simulated in <1 hour, are shown in Figure 10. These simulations can then be extended to local variability aware HSPICE simulations of any given library cell.

It is important to note that this TCAD-SPICE tool flow is unique in its ability to maintain all correlation effects from global process variations to local device variability all the way down to standard cell simulations.

IV. CONCLUSION

Scaling of future semiconductor technologies will rely increasingly on new device architectures and materials to provide the required power, performance and area benefit expected of new technology nodes. Given the large number of device architectures and material options available, semiconductor manufacturers face a significant challenge in evaluating all available options and selecting the best process/design since it is too time-consuming and costly to evaluate each option with hardware. To address this challenge, TCAD has evolved from primarily simulating single transistors to becoming an integral part of design-technology co-optimization. This "Shift Left" strategy requires not only advanced physics to be captured in TCAD tools, but also tool connectivity and automation to handle the complexity of modeling the interactions and impact between design and process. The two TCAD-Based DTCO examples shown in this paper illustrate the value of TCAD in early stages of technology development for guiding decisions and enabling selection of technology options through early process, design rule, and library development.

REFERENCES

[1] Jeff Wu, "Expanding Role of Predictive TCAD in Advanced Technology Development," Proc. of SISPAD, p. 167 - 171, 2013.
[2] J. P. Colinge, "Multigate Transistor: Pushing Moore's Law to the Limit," Proc. of SISPAD, pp. 313 - 316, 2014.
[3] T. Ma et al., "TCAD: Present State and Future Challenges," IEDM Proceedings, pp. 367 – 370, 2010.
[4] Keun-Ho Lee, "Challenges and Responses for Virtual Silicon," Proc. of SISPAD, pp. 80 - 83, 2015.
[5] Victor Moroz et al., "Power-Performance-Area Engineering of 5nm Nanowire Library Cells," Proc. of SISPAD, pp. 433 – 436, 2015.
[6] Xingsheng Wang et al., "FinFET Centric Variability-Aware Compact Model Extraction and Generation Technology Supporting DTCO", IEEE Transactions on Electron Devices, Vol. 62, No. 10, pp. 3139-3146, 2015
[7] Asen Asenov et al., "Variability Aware Simulation Based Design-Technology Co-Optimization (DTCO) Flow in 14nm FinFET/SRAM Cooptimization", IEEE Transactions on Electron Devices, Vol. 62, no. 6, pp. 1682-1690, 2015.
[8] F. Adamu-Lema et al., "Comprehensive 'Atomistic' Simulation of Statistical Variability and Reliability in 14 nm Generation FinFETs", Proc. of SISPAD, pp. 157 – 160, 2015.
[9] U.S. Patent Application No. 15/149,994, filed May 9, 2016, "Parameter Generation for Modeling of Process-Induced Semiconductor Device Variation.

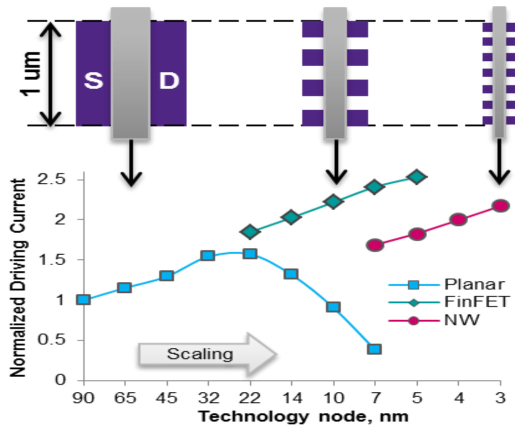

Figure 1. Transistor strength evolution per 1µm layout footprint and fixed off-state current and power supply.

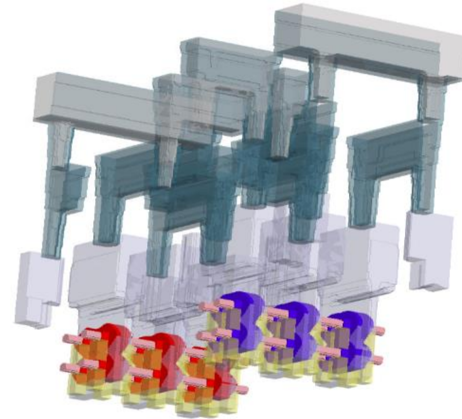

Figure 4. A 2-NAND logic cell with 5nm design rules and two level stacked lateral nanowires.

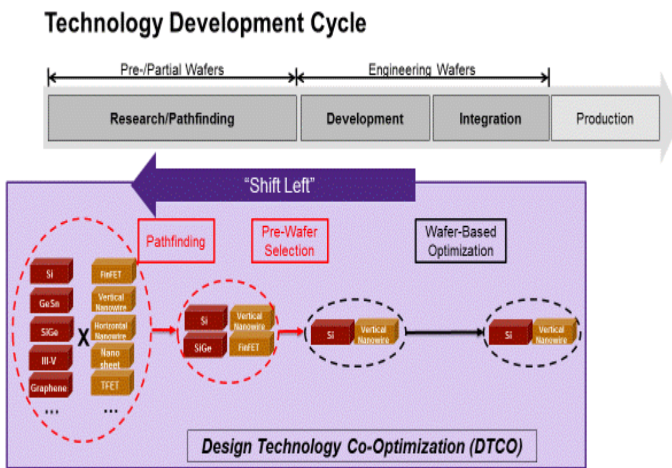

Figure 2. TCAD "Shift Left" to support the selection of technology options with design-level criteria in the early, pre-wafer stages of technology development.

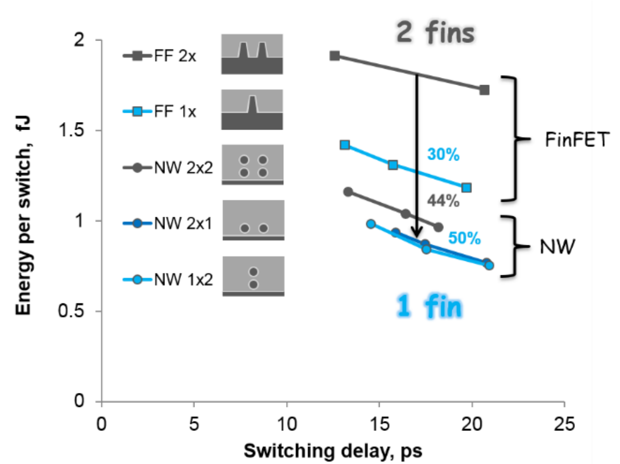

Figure 5. Summary of benchmarking different transistor architectures for 2-NAND logic cell with 5nm design rules and a load of fan-out of 2 and a 70 metal pitches long BEOL wire. Power savings are shown in % w.r.t. reference structure where 2 fins are connected in parallel.

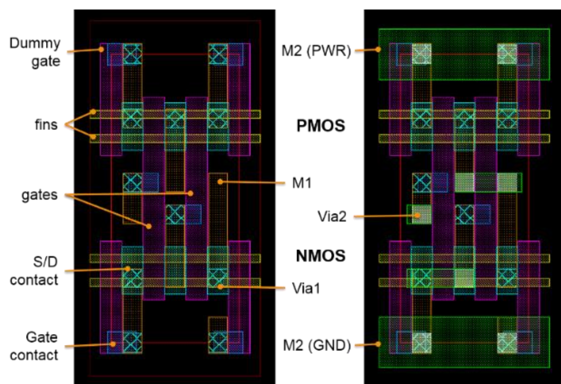

Figure 3. Layout of a 9 track tall 2-input NAND logic cell with 5nm design rules, 2 NMOS and 2 PMOS fins.

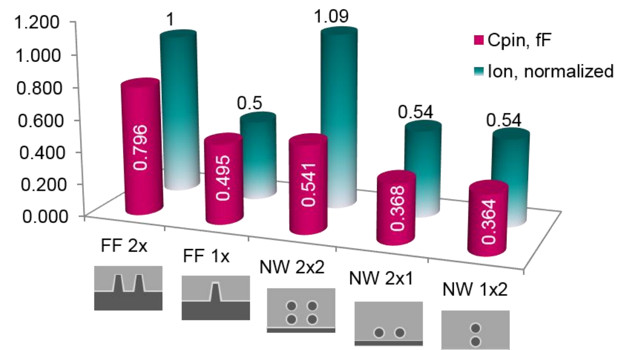

Figure 6. Transistor strength normalized to the 2-fin option and pin capacitance listed for all five transistor architectures.

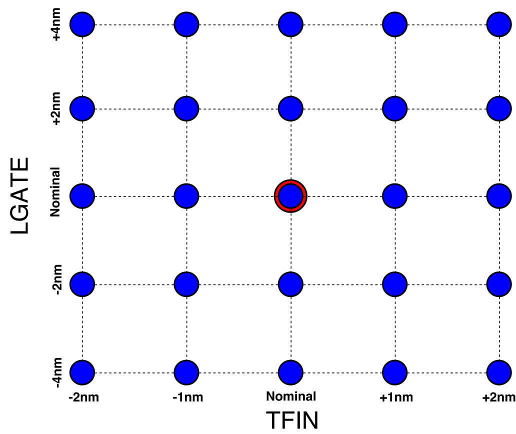

Figure 7: DoE space in this demonstration flow – for both nMOS and pMOS. The red circle denotes the nominal transistor design point. For the purpose of this demonstration we will only consider 2 process axes – fin thickness (TFin) and gate length (Lgate).

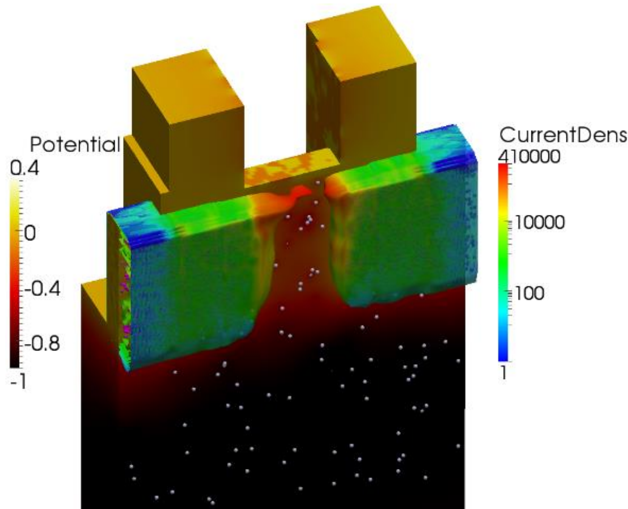

Figure 8. An example of an 'atomistic' 14nm FinFET simulation with RDF, GER, FER and MGG as local variability sources. The potential is in Volts and the current density is in A/cm^2. Half of the fin is visualized and colored by potential (showing the different grains in the gate). Electron contours show a current path though the channel as well as the dopant configuration.

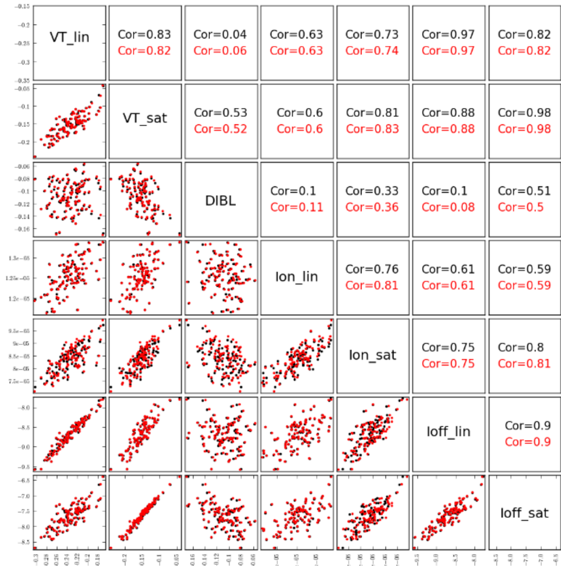

Figure 9. Comparison of the distributions and correlations of key transistor figures of merit obtained from the TCAD simulations and from the extracted statistical spice compact models. The black points represent the simulated TCAD and the red points represent the extracted compact models.

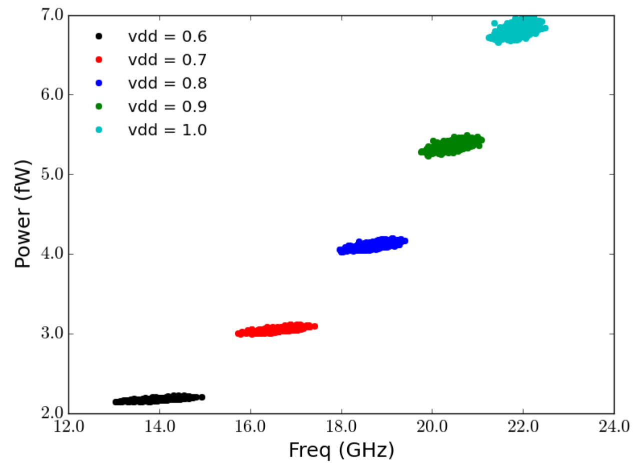

Figure 10. Ring Oscillator results for an nMOS/pMOS pair in the DoE. This example is a 7 stage RO with 0.5fF capacitive load at multiple power supply levels.

Simulation of Semiconductor Processes and Devices 2016
Edited by E. Bär, J. Lorenz, and P. Pichler