TCAD challenges and opportunities to find a feasible device architecture for sub-3nm scaling

Uihui Kwon*, Yonghee Park, Yoon-Suk Kim, Jaehyun Yoo, Dae Sin Kim
CSE Team, DIT Center, Samsung Electronics Corp. Ltd., Hwasung-si, Gyeonggi-do, Korea
*e-mail: uihui.kwon@samsung.com

Abstract

With the aggressive scaling of MOSFET devices below 3nm, the role of TCAD in selecting a feasible device architecture for next node has become extremely important. There is an enormous opportunity cost for each choice, so the pros and cons of each option must be identified through seamless pre-validation using TCAD. Therefore, it is important to understand which TCAD solutions are necessary to validate the architecture candidate in rigorous way. This paper describes which TCAD solutions are important in atomic/device/standard cell/block-chip level for next generation logic pathfinding from the perspective view of a semiconductor manufacturing industry.

1. INTRODUCTION

The logic generation from 65nm to 32/28nm based on planar transistor, was a performance booster-driven scaling era such as stressor/HKMG, whereas from 22nm to current 5nm based on FinFET could be called a DTCO-driven scaling era, and sub-3nm node will be an architecture-driven scaling era, which is visualized in Fig 1.[1]

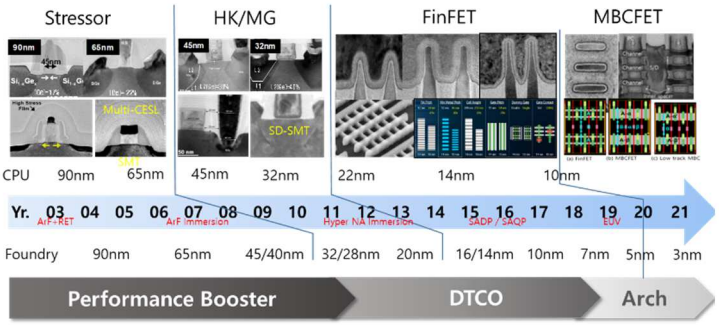


Fig 1. Logic scaling trend

In the DTCO-driven FinFET era, various design elements such as single diffusion break (SDB), contact over active gate (COAG), and a new middle of line (MOL) scheme have been introduced in order to reduce the cell size and to minimize the Reff/Ceff in standard cell (STC) level. At the end of the lifetime of FinFET, which has played the role of major device architecture, the scaling scenarios we can take are largely categorized into three options.

The first scenario is to introduce new materials such as high mobility channel or ferroelectric gate stack while maintaining the device architecture as FinFET. Various attempts have been made so far, such as SiGe / Ge / III-V channel materials and ferroelectric gate material, which are still on-going. SiGe, one of the oldest candidates for new channel material, has the limitation that its bulk mobility is worse than that of Si due to alloy scattering [2]. However, it has finally been productized as a multi-Vth solution.[3] On the other hand, Ge is certainly a high mobility material, but it's too small bandgap causes large BTBT leakage in SD region.[4] As for III-V materials in ultra-thin body (UTB), the mobility degradation by carrier spill-over into L-valley and density of state (DOS) bottleneck are still issues.[5]

The second scenario is to keep the material as Si and change the device architecture to GAA structure such as lateral or vertical nanosheet. Lateral schemes are highly optimized because DTCO has been progressed for several generations from 22nm to 5nm, on the other hand, vertical schemes such as VFET still have a lot of room for further optimization. In recent years, with the spread of Si stacking technology, CFETs that construct CMOS by stacking N and P are attracting wide attention. [6]

The last scenario is to change both the material and the device architecture simultaneously, which is the riskiest option. However, if the new architecture can compensate for the shortcomings of the new material, it might become an unexpected game changer. The following Table.1 summarizes the TCAD solutions which are important in order to hedge the risk of each option.

Table 1. Key TCAD solutions for each option

Categories	New Material	Arch Change
Atomic	Gox/ Silicide	Multi-WF Dipole
Device	Injection velocity	2D Quantum
Std. Cell	Compact model	Intelligent DTCO
Block/Chip	Self-heating	STCO (PDN)

2. TCAD Challenges & Opportunities

To overcome the daunting challenges of logic technology definition below 3 nm node, our TCAD solutions in atomic level / device level / Std. cell level / IP block & chip level should be reinforced with new methodologies and ready for real application at least 1year in advance.

978-1-6654-0685-7/21/$31.00 ©2021 IEEE

(1) Atomic Level

Regardless the scaling scenarios of changing channel material or transistor architecture, it is mandatory to discover a new performance booster lowering resistance (R) and capacitance (C) through a full atomistic simulation capable of dealing with all major physical effects in atomistic level such as bottom-up approach as shown in Fig 2. [7,8]

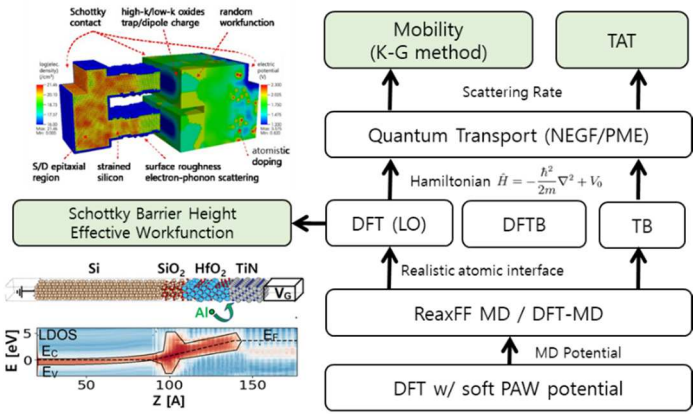


Fig 2. A full atomic simulation of bottom-up approach

Especially, with the advent of SiGe FinFET [3] and nanosheet devices utilizing SiGe epi-layer in channel [11], the possibility that carrier mobility may deteriorate by SiGeO formed at the gate oxide interface has increased. In this a case, even though the geometrical roughness measured by AFM looks good, the electrical roughness can be hopelessly bad. Fig 3 shows how to distinguish between the electrical roughness of Si/SiO_2 and that of Si/SiGeO through full atomistic DFT simulation. [9]

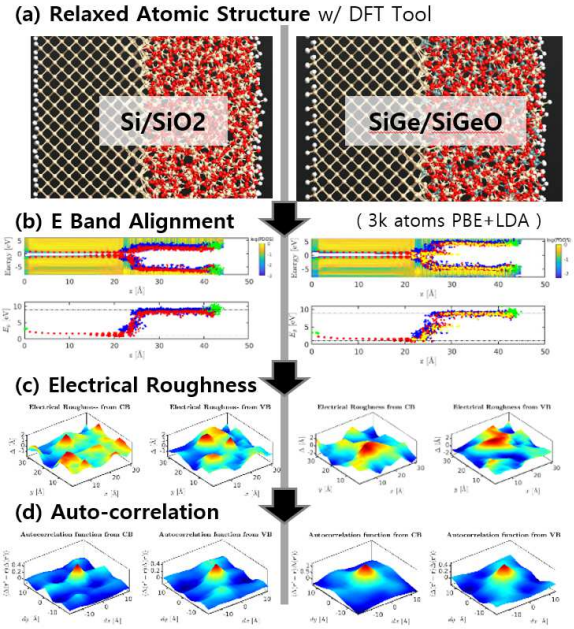


Fig 3. Full atomistic extraction of electrical roughness

Moreover, in the case of GAA devices, as the space between nanosheets, which should be filled with gate oxide and workfunction (WF) metal, becomes narrower, so it is crucial to simulate gate WF precisely for the full gate stack with considering all the WF-affecting agents like Al, oxygen, and dipole agents with bottom-up approach. [8]

(2) Device Level

When channel material or device architecture is changed, the probability of defect generation and the efficiency of the SD embedded stressor like eSiGe change. It is reported that eSiGe stress can be degraded by stacking fault generation in epi SD region in FinFET as shown in Fig 4 (a).[10] Even in the stacked nanosheet device having separated Si seed surfaces for epi-growth [11], it is essential to minimize the strain loss caused by high dimensional defects like stacking fault and grain boundary with proper dislocation stress field model as shown in Fig.4(b). [12]

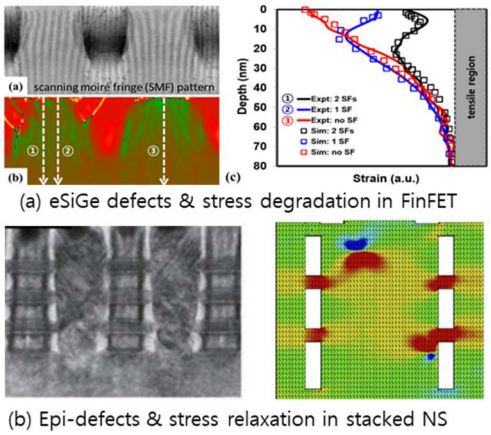


Fig 4. Modeling of eSiGe epi-stress degradation by defect

With the advent of a full-fledged GAA device, gate length (Lg) scaling has become mandatory within the range of minimum short channel effect deterioration. To decide the optimal Lg, it is important to calculate the available states under 2D quantum confinement in channel cross-section and the quasi-ballistic carrier transport along SD direction rigorously, solving a multi-subband Boltzmann transport equation (MS-BTE) as shown in Fig 5. [13]

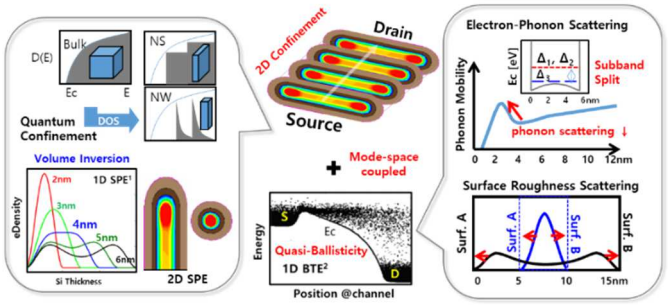


Fig 5. MS-BTE solver for advanced transport in GAA

Particularly, it is crucial to optimize the carrier injection at virtual source (VS) using this as shown in Fig 6, where the apparent mobility (μ_{app}) and injection velocity (v_{inj}) are defined as below. If you can extract the apparent mobility and injection velocity of your Si HW based on these virtual source model and compare it, You can easily figure out the remaining room for further optimization depending on device.

$$\mu_{app} \sim \frac{L_{eff} \cdot W}{R_{ch} \cdot C_G (V_{DD} - V_{TLin})} \sim \frac{L_{eff} \cdot (2H_{Fin})}{R_{ch} \cdot C_G (V_{DD} - V_{TLin})}$$

$$v_{inj} \sim \frac{I_{dSat} \cdot W}{C_G (V_{DD} - V_{TSat})} \sim \frac{I_{eff\_T} \cdot (DC/100) \cdot (2H_{Fin})}{C_G (V_{DD} - V_{TLin} + DIBL)}$$

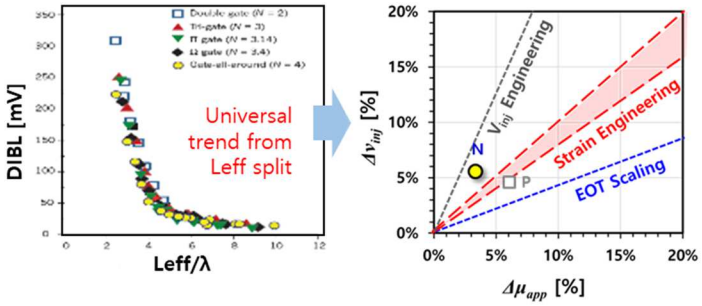
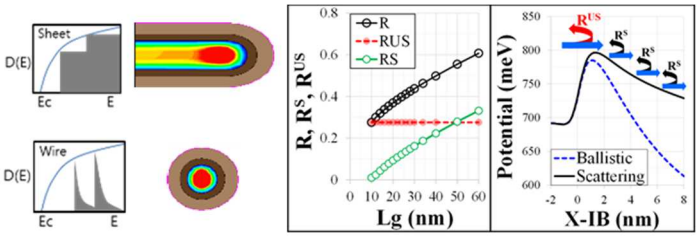


Fig 6. Engineering of carrier injection at virtual source

In addition, device compact model for SPICE simulation should be improved to properly reflect these physics, the 2D quantum confinement and quasi-ballisticity, what we call a physical compact model, which can provide interesting new features, e.g. predictive forms with few parameters. In this regard, virtual-source (VS) model offers an efficient platform, as it covers the quasi-ballistic nature of transport. The core model of VS can be further improved to entail essential details. For example, as shown in Fig 7(a), 2D confinement and corresponding density of state should be included for advanced nodes. Moreover, unscalable reflections becomes increasingly important for short devices, refer to Fig 7(b) [14,15].



(a) 2D Quantum effect

(b) Decoupling the reflection (R) components into scalable (R_S) and unscalable (R_{US}) one

Figure 7. Reinforced physics for stacked GAA devices

Including all these physics, we obtain precise Lg-sensitivity for injection- and scattering-current, where surface roughness & phonon scattering are considered, as shown in Fig 8. The accuracy of injection-current implies the correct barrier height and electrostatics, whereas fitting the scattering-current indicates the quality of transmission rate below. In this way, the electrostatics and transport metrics are accurately captured over the entire bias range.

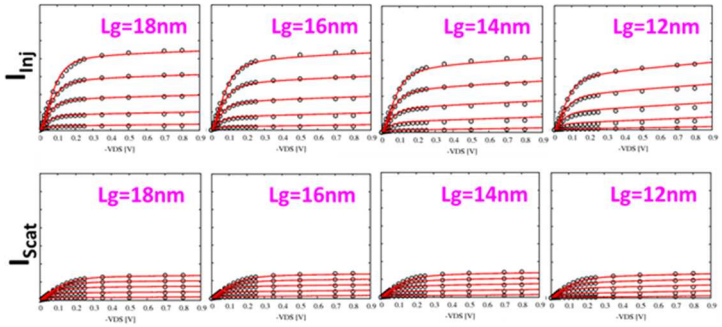
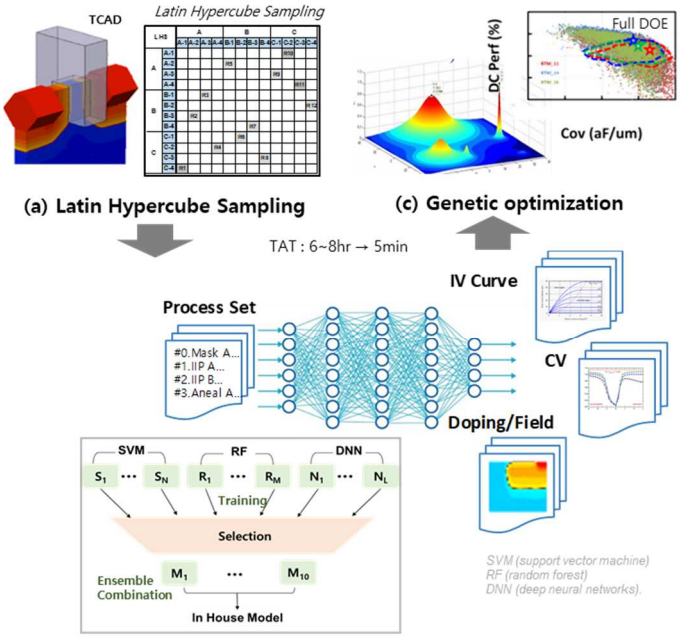


Fig 8. Injection (I_{Inj}) and scattering current (I_{Scat}) by MSBTE (symbols) and compact model (lines) at fixed Ioff

Finally, in order to timely provide the optimal process & device condition for various devices in foundry business, the process/device TCAD simulation time, typically 4 ~ 8 hours per a job, is also a big burden. Therefore, it is important to secure a real-time TCAD (RTT) methodology that learns TCAD simulation data to create a high-quality surrogate model and uses it to find the optimal process condition & device geometry through massive full DOE. Fig 9 shows our approach to embody RTT and the machine learning method used for device optimization. [16,17]



(b) Response surface model with ensemble combination

Fig 9. Machine-learning based device optimization

For a typical legacy device, the simulation speed of RTT is 100k~500k times faster than conventional TCAD, and the process

development time is reduced by 2~3 weeks. The whole procedure described here, 1) data generation by TCAD simulation, 2) training model (accuracy >93%), and 3) genetic optimization with surrogate model, is fully automated to realize RTT.

(3) Standard Cell Level

The importance of the holistic optimization of standard cell (STC), what we call design technology co-optimization (DTCO), cannot be exaggerated. Fig. 10 shows the DTCO-driven scaling conducted for FinFET so far [18]; (a) changing the double diffusion break (DDB) to single diffusion break (SDB) to reduce the STC width, (b) reducing the distance to the power rail or the space between N/P active region to scale cell height, and (c) securing CB-to-CB margin with a contact on active gate (COAG).

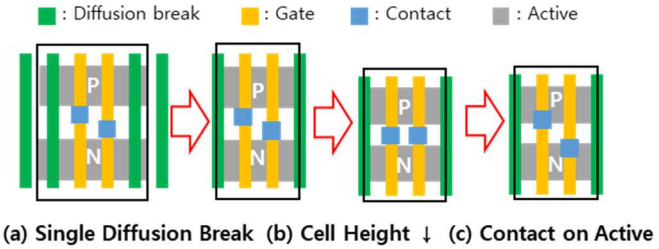


Fig 10. DTCO-driven scaling used for FinFET

Different from the conventional lateral schemes such as FinFET and MBCFET, vertical scheme like VFET still has more room for DTCO. Fig 11 shows the typical example of DTCO flow for vertical scheme.

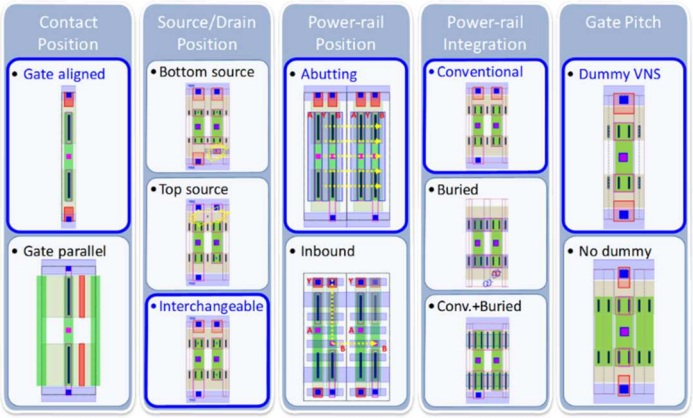


Fig 11. DTCO flow for vertical scheme

The cell size scaling usually leads to an increase of Reff/Ceff due to the decreased critical dimensions, which leads to the degradation of performance and power. Moreover, it leads to process yield drop by reducing process margins such as short & overlap. Moreover, with the introduction of new device architectures such as GAA/CFET, the number of design parameters affecting cell performance has increased a lot, so it has become extremely important to check its performance-power-area-yield (PPAY) impact quickly through 3D process emulation in advance as shown in Fig 12.

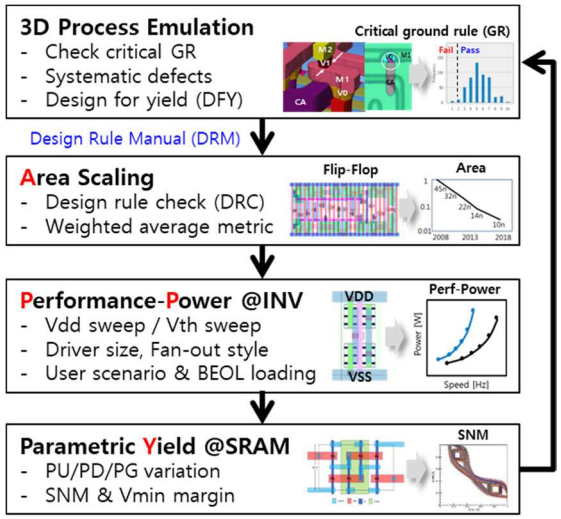


Fig 12. PPAY optimization with iDTCO

Especially, ML-based intelligent DTCO (iDTCO) technology that can optimize multiple Y (speed, power, and yield) for multiple X factors (Layout, PA, MTS …) is a crucial one [1], which is described in Fig 13.

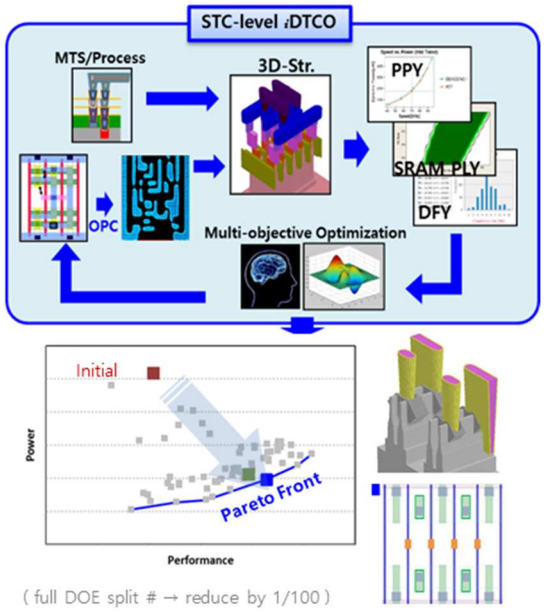


Fig 13. Intelligent DTCO platform

(4) IP Block/Chip - Level

With the introduction of 1-fin cell to minimize dynamic power, it has become crucial to compare the speed & power trade-off of each architecture candidate in block level for fair comparison. In addition, as the difficulty of co-integration of EG devices for I/O

increases rapidly, people are getting more interest on BEOL power gating devices and backside power distribution network. In particular, with the emergence of 3D VLSI concepts such as stacked cell such as CFET or Monolithic 3D, system technology co-optimization has become essential.

Traditionally, local layout effect (LLE) was invented to reflect well proximity effect (WPE), the effect of ion implantation shadowing at the well boundary, and STI stress effect according to length of diffusion size to process design kit. [19] However, with the advent of HK/MG process, many WF-caused LLEs such as metal gate boundary (MGB) have become dominant. Moreover, the recent understanding that single diffusion break (SDB) and double diffusion break (DDB) apply different STI stress brought the birth of a mixed diffusion break (MDB) which uses SDB for PMOS and DDB for NMOS. [18] Including CT-cut affecting both ILD stress and WF fluctuation, advanced LLE modeling that can distinguish stress-caused LLE, (SASB: length of diffusion, RXH: RX horizontal spacing, PCP: poly contact pitch, CT: gate cut) and WF-caused LLE (WPE: well proximity effect, MGB: metal gate boundary, RXV: RX vertical spacing, PPR: PC past RX) is crucial as shown in Fig 14. [19,20]

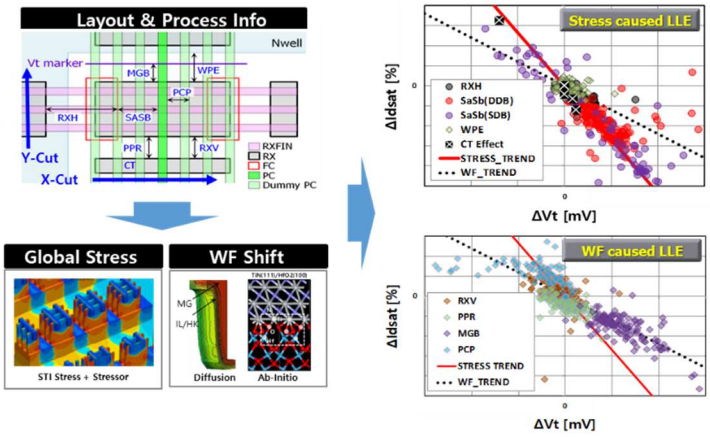
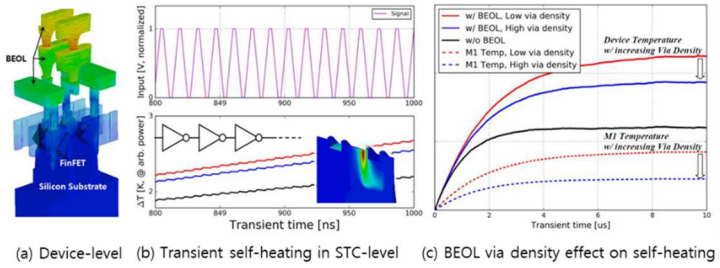


Fig 14. Classification of local layout effect (LLE)

With the advent of materials and architectures that are disadvantageous to self-heating such as SiGe channels [21] and CFETs, the pre-evaluation of the electro-thermal reliability in block/chip level has become a mandatory sign-off flow.



(a) Device-level (b) Transient self-heating in STC-level (c) BEOL via density effect on self-heating

Fig 15. Multi-scale self-heating modeling

So it is crucial to secure a multi-scale electro-thermal simulation bridging from carrier energy transport in transistor level to thermal circuit modeling in IP block level and to predict its lifetime. Fig 15 shows the examples of multi-scale self-heating modeling to assess BEOL design impact on device temperature. [22]

Finally, with the growing market of power & infotainment devices for automotive application, a fast and accurate tool for soft-error rate (SER) has become indispensable. Fig.16 shows the schematic flow of our in-house SER simulator based on a physical charge collection model. [23]

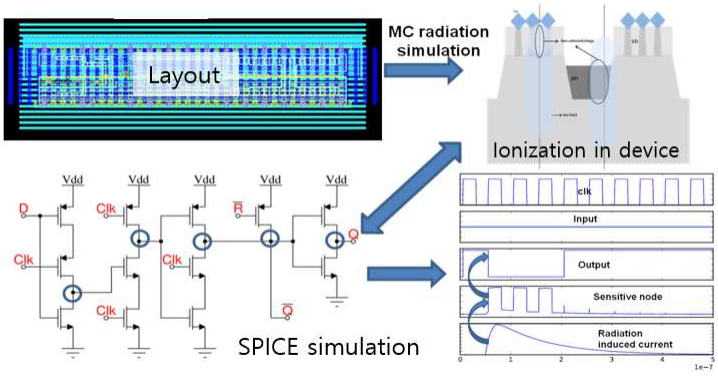


Fig 16. Soft error rate (SER) simulation flow

3. CONCLUSION

To summarize, (1) atomic level: a full atomistic simulation based on bottom up approach is important to discover a new performance booster, (2) device level: advanced stress & virtual source engineering based on MS-BTE solver and its speed-up with real-time TCAD (RTT) solution are crucial, (3) STC level: PPAY should be co-optimized in the early stage of technology definition with an accurate, computationally inexpensive DTCO framework, our iDTCO framework could speeds up daily PPAY analysis by 5~10 times with good accuracy. (4) IP block & chip level: advanced LLE modeling is mandatory for holistic optimization in block level and fast-but-accurate self-heating & soft error rate modeling are crucial for server & automotive applications.

In order for the aforementioned TCAD solutions to be fast enough so to be applied to product development in a timely manner, it is also important to continuously develop computational acceleration technology that utilizes the state-of-the-art GPGPU technology and ML algorithms. These speed-up solutions should be supported by physical simulations in device and material level facilitated with domain knowledge, so multi-scale device & process modeling connecting from atomistic scale to standard cell scale is indispensable.

REFERENCES

[1] Uihui Kwon, et. al., “Intelligent DTCO (iDTCO) for next generation logic path-finding”, SISPAD 2018

[2] Changwook Jeong, et. al., “Physical understanding of alloy scattering in SiGe channel for high-performance strained pFETs”, IEDM 2013

[3] G. Yeap, et. al., “5nm CMOS Production Technology Platform featuring full-fledged EUV, and High Mobility Channel FinFETs with densest 0.021µm2 SRAM cells for Mobile SoC and High Performance Computing Applications“, IEDM 2019

[4] Siddhartha S. Dhar, et. al., “Impact of BTBT, stress and interface charge on optimum Ge in SiGe pMOS for low power applications”, SISPAD 2016

[5] R Kim, et. al., “Comprehensive n-and pMOSFET channel material benchmarking and analysis of CMOS performance metrics considering quantum transport and carrier scattering effects”, J EDS 8, 505-523 (2020)

[6] P. Schuddinck, et. al., “Device-, Circuit- & Block-level evaluation of CFET in a 4 track library”, VLSI 2019

[7] Hong-Hyun Park, et. al., “Toward more realistic NEGF simulations of vertically stacked multiple SiNW FETs”, SISPAD 2018

[8] H. Ilatikhameneh, et. al., “Effective work-function tuning of TiN/HfO 2 /SiO 2 gate-stack; a density functional tight binding study”, SISPAD 2019

[9] K. Vuttivorakulchai, et. al., “First-principle Extraction of Surface Roughness in Si/Oxide Interfaces”, SISPAD 2021

[10] Uihui Kwon et. al., “Progress in dislocation stress field model and its applications”, SISPAD 2019

[11] N. Loubet, et. al., “Stacked Nanosheet Gate-All-Around Transistor to Enable Scaling Beyond FinFET”, VLSI 2017

[12] Chihak Ahn, et. al., ” A finite element method to simulate dislocation stress: A general numerical solution for inclusion problems”, AIP Advances 10, 015111 (2020)

[13] Seonghoon Jin, et.al., “Coupled Drift-Diffusion (DD) and Multi-Subband Boltzmann Transport Equation (MSBTE) Solver for 3D Multi-Gate Transistors”, SISPAD 2013

[14] K. Natori, et. al., “Anomalous degradation of low field mobility in short-channel metal-oxide-semiconductor field-effect transistors,” J. Appl. Phys., vol. 118, no. 23, Dec. 2015

[15] M. A. Pourghaderi, et. al., "Ballisticity Saturation by Unscalable Reflections," Electron Device Letters, vol. 41, no. 7, pp. 969-972, July 2020

[16] Sanghoon Myung, et. al., “Real-Time TCAD: a new paradigm for TCAD in the artificial intelligence era”, SISPAD 2020

[17] Jaehyun Yoo, et. al., “Machine-Learning based TCAD Optimization Method for Next Generation BCD Process Development”, ISPSD 2021

[18] ES Jung et.al., Samsung Foundry Forum 2018

[19] Choongmok Lee, et. al., “Layout-induced stress effects on the performance and variation of FinFETs”, SISPAD 2015

[20] Pei Zhao et. al., “Influence of stress induced CT local layout effect (LLE) on 14nm FinFET “, VLSI 2018

[21] Anh-Tuan Pham, et. al., “Simulations of Self-Heating Effects in SiGe pFinFETs Based on Self-Consistent Solution of Carrier/Phonon BTE Coupled System”, SISPAD 2018

[22] Jaehee Choi, et. al., “Impact of BEOL Design on Self-heating and Reliability in Highly-scaled FinFETs”, SISPAD 2019

[23] Udit Monga, et. al., “Charge-collection Modeling for SER Simulation in FinFETs”, SISPAD 2016