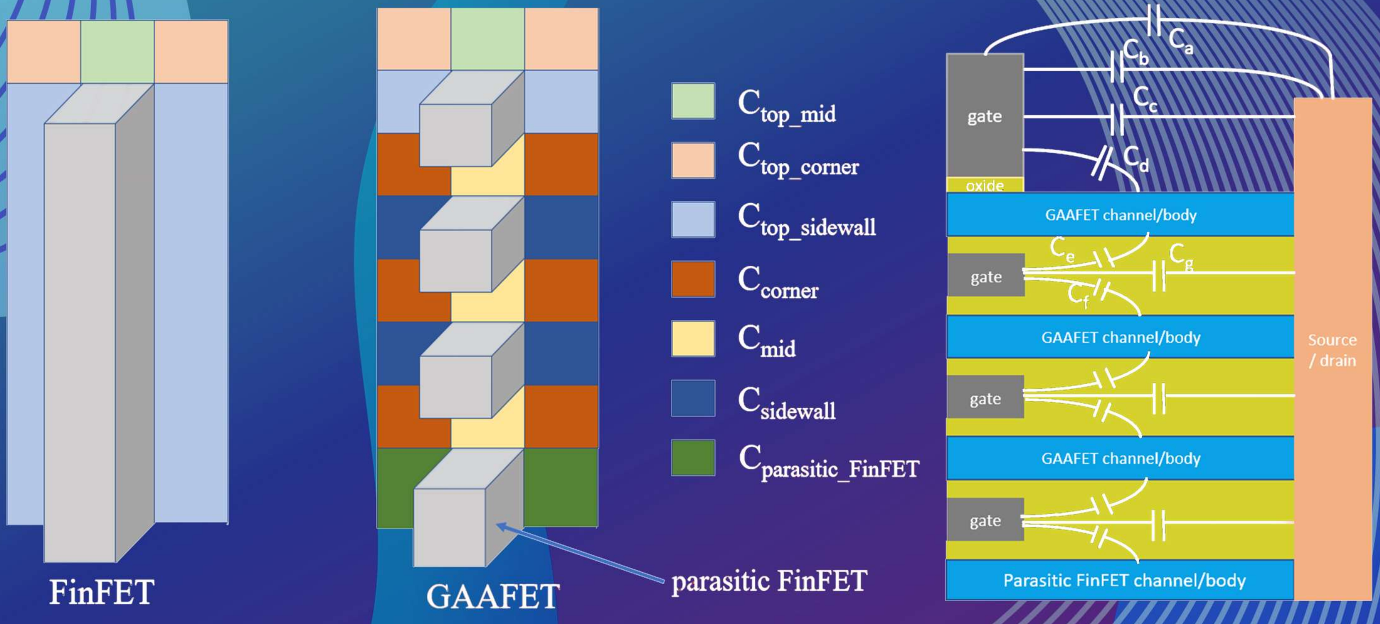
JOMM

Journal of Microelectronic Manufacturing

Online ISSN: 2578-3769 Volume 3, Issue 4, Dec. 2020



Graphical representation of the parasitic capacitance components for FinFETs and GAAFETs.

Author: Avirup Dasgupta, Chenming Hu

J. Microelectron. Manuf. 3, 20030402 (2020)

www.jommpublish.org

JOURNAL OF MICROELECTRONIC MANUFACTURING

VOLUME 3, ISSUE 4 DECEMBER 2020

TABLE OF CONTENTS

J. Microelectron. Manuf. **3**, 20030401 (2020)
doi: 10.33079/jomm.20030401

Guest Editorial: Special Issue on CAD Technologies

Moore's Law has been a major driving force for the exponential growth in the semiconductor industry for nearly the past six decades, regularly doubling the number of transistors on a chip and increasing performance twofold approximately every two years. In recent years, there have been position papers predicting the imminent demise of Moore's Law as many physical limits are reached, mainly as a result of decades of aggressive technology scaling, lithography, non-scaling interconnection with technology nodes, random and systematic process variabilities, and more. Yet, the global semiconductor market size reached USD 513 billion in 2019 and is projected to reach USD 727 billion by 2027 on a CAGR of 4.7%. This projected growth can be attributed to the increasing consumption of consumer electronic devices globally as well as the emergence of big data, artificial intelligence, machine learning, internet of things, and 5G that provide tremendous new opportunities to the market growth. To sustain the semiconductor industry's continued growth, however, the continuation of Moore's Law or its variants must be realized which in turn requires unprecedented parallel R&D efforts on novel transistor architecture, new materials, efficient computational lithography, design and technology co-optimization, advanced packaging, and effective manufacturing yield improvement, etc.

Facing the ever more complex and challenging development of process technologies in the nanometer regime, advanced computer-aided design (CAD) technologies have become indispensable enablers for early pathfinding, transistor and backend definition and optimization, design & technology co-optimization for performance-power-area and reducing the risk in re-design, novel material exploration, lithography and OPC development, defect detection and yield improvement, etc. It is our great pleasure to present this special issue of the Journal of Microelectronic Manufacturing on "CAD Technologies Enabling Advanced Process Technology Development and Product Design." This issue contains nine invited papers authored by distinguished scholars and researchers from leading universities, research institutes, and the industry. The topics covered include: (a) an industry-standard physical Spice model for FinFET to Gate-All-Around FET; (b) three Technology CAD (TCAD) device simulation papers discussing the Scharfetter-Gummel discretization scheme in solving the drift-diffusion transport model, an advanced open-source TCAD simulation platform, and a 1st principle-based TCAD simulation applied to the design of tunnel FET; (c) two papers on computational lithography and OPC utilizing machine learning; (d) one paper on TCAD-based methodology to enable design-technology co-optimization of advanced semiconductor memories including a multi-stage simulation flow to study the device-to-circuit performance in presence of statistical and process variability; (e) one paper on applications involving a complex set of material modeling tools and methodologies and share a perspective of the future of the area; and (f) one paper on a comprehensive pattern centric platform for process technology development and manufacturing.

We would like to express our sincerest gratitude to the authors for their gracious and insightful responses to our invitation to contribute to this special issue of JoMM. We sincerely appreciate their time, effort and support. Also, we would like to thank all the reviewers for their meticulous review and expert suggestions.

Shiuh-Wuu Lee *Guest Editor*

Dr. Shiuh-Wuu Lee worked in the semiconductor industry for 37 years with 20 years at Intel Corporation as Intel Fellow and director, 8 years at AT&T Bell Laboratories as member of technical staff, 6 years at SMIC as executive vice president and general manager of Technology Development, 1.5 years at Synopsys as special consultant to its Silicon Engineering Group and 1 year at Zhejiang Laboratory as director of its chip design center. Elected to China's National Thousand Experts Program in 2013, received PhD degree in electrical & computer engineering from the University of Michigan in 1980, published 61 technical journal and conference papers, and delivered over 45 keynote and invited speeches at major industry forums and international technical conferences.

J. Microelectron. Manuf. **3**, 20030402 (2020)
doi: 10.33079/jomm.20030402

BSIM-CMG Compact Model for IC CAD: from FinFET to Gate-All-Around FET Technology

Avirup Dasgupta [*] and Chenming Hu [**]

Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA.

Abstract: We discuss the BSIM-CMG compact model for SPICE simulations of any common multi-gate (CMG) device. This is an industry standard model which has been used extensively for FinFETs IC design and simulation, and has now been extended to accurately model gate-all-around FET (GAAFET). We present the core framework of BSIM-CMG and discuss the latest updates that capture various physical phenomena originating from the quantum confinement of electrons by the small cross section of the GAAFET channel. Special attention is paid to providing suitable model parameters that can be adjusted using software tools to match the model with manufactured transistors very accurately. Furthermore, the model's speed allows the use of Monte Carlo circuit simulation to account for random device variations encountered in manufacturing. This model is the industry standard compact model for GAAFETs and will help bridge the wide divide between GAA IC manufacturing and design, starting at 3nm/2nm technologies.

Keywords: Gate-all-around, GAAFET, FinFET, BSIM, BSIM-CMG, Compact model, Quantum, Nanosheet,3D, Transistor.

1. Introduction

Semiconductor devices have continuously improved over the past few decades in terms of density, performance and power consumption. This has been brought about mainly by scaling of transistors [1,2]. One of the most significant events for the semiconductor industry was the shift to FinFETs [1,2]. The design of these devices allows a gate on three sides of the channel resulting in greater gate control. This is of utmost importance to negate the side-effects of scaling (short channel effects). Moreover, the 3D vertical structure reduces the area requirement and allows further increase of the circuit density.

To continue scaling further, we require even greater gate control. The next logical step after FinFETs is to have gate on all sides of the channel; giving rise to the Gate-All-Around FET (GAAFET), as shown in Figure 1 [3-6]. Several companies have recently announced the use of GAAFETs for production design [3-10]. This device not only provides excellent gate control, but also utilizes a vertical structure with multiple channels per fin to reduce the footprint even further [7].

Designing circuits with such devices requires a compact model for SPICE simulators. The device model is a set of equations that describe the device behavior and can be evaluated very fast so that very large circuits can be simulated while being able to reproduce the very complex transistor characteristics accurately. It needs to be accurate to avoid expensive re-designs, very fast to enable timely simulation of large circuits as well as robust to ensure convergence for a wide range of complex circuits and simulation conditions [2]. BSIM-CMG is the industry standard models for common-multi-gate (CMG) devices like FinFETs and GAAFETs. The model can accurately simulate double gate, triple gate, quadruple gate and gate-all-around structures of any geometry including commercial FinFET and GAAFET devices.

In this paper, we will provide an overview of the BSIM-CMG compact model with special emphasis on GAAFETs.

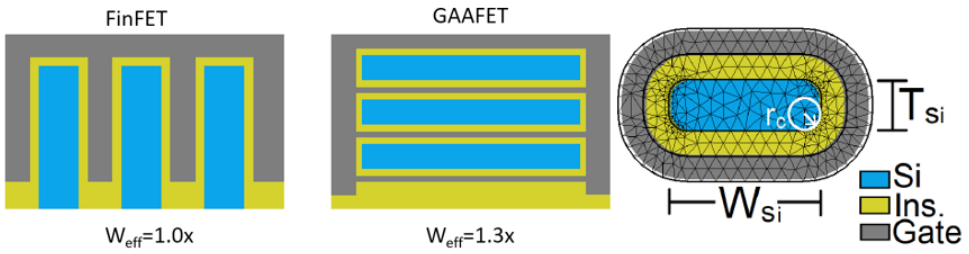
2. BSIM-CMG Core Framework

The BSIM-CMG model is a compact (SPICE) model for common-multi-gate devices [2]. It is based on a core model which calculates the device electrostatics and transport using a long-channel assumption. Physical effects like short channel effects, leakage currents, non-quasi-static effects, noise etc., are added on top of the core model as demonstrated in Figure 2.

The core electrostatics is based on the Poisson

[*] Address all correspondence to Avirup Dasgupta, E-mail: avirup@berkeley.edu

[**] Address all correspondence to Chenming Hu, E-mail: hu@eecs.berkeley.edu



(a) (b)

Figure 1. (a) Graphical representation of FinFET and GAAFET. The GAAFET structure. (b) GAAFET cross-section used for band-structure TCAD simulations; illustrating the width (W_{Si}), thickness (T_{Si}) and the corner radius (r_c).

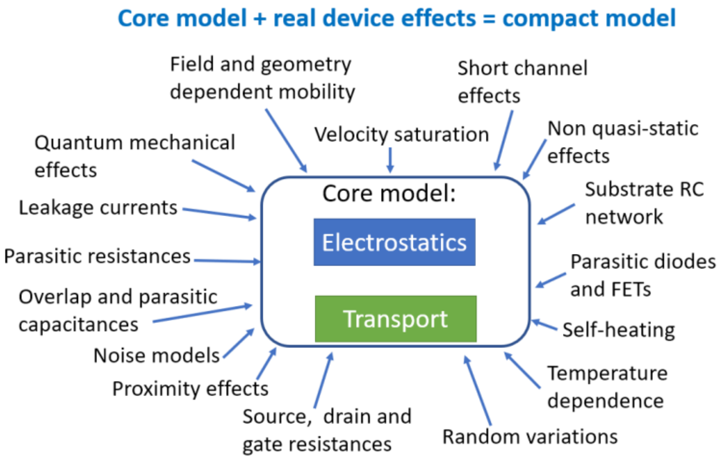


Figure 2. An illustration of the BSIM-CMG compact model framework.

equation with several approximations. The core equation can be given as [2,15]

$$v_G - v_0 - v_{ch} - \Delta q_{dep} = \underbrace{-q_m}_{strong\ inversion} + \underbrace{\ln(-q_m)}_{weak\ inversion} + \ln(\underbrace{\frac{q_t^2}{e^{q_t} - q_t - 1}}_{moderate\ inversion}), \quad (1)$$

where v_0 and q_t represent

$$v_0 = v_{FB} - q_{dep} - \ln(\frac{2qn_i^2 A_{ch}}{v_T C_{ins} N_{ch}}), \quad (2)$$

$$q_t = (q_m - q_{dep}) r_N. \quad (3)$$

In these equations, q is the electronic charge, n_i is the intrinsic carrier concentration, v_T is the thermal voltage, $v_G = V_G / v_T$ is the normalized gate voltage, and $v_{ch} = V_{ch} / v_T$ is the normalized channel voltage. Also, $q_m = Q_m / v_T C_{ins}$ with Q_m denoting the mobile charge density and q_{dep} = $-qN_{ch} A_{ch} / v_T C_{ins}$ is the normalized depletion charge density. The term r_N is defined as $r_N = A_{Fin} C_{ins} / \epsilon_{ch} W_{eff}$, where ϵ_{ch} is the permittivity of the channel and A_{Fin} is the area of the fin. The term Δq_{dep} accounts for the effect of body bias for FinFETs fabricated over bulk substrates. This term is defined as [15]

$$\Delta q_{dep} = -\frac{\gamma}{2v_T}\left(\sqrt{2v_T \ln\left(\frac{N_{ch}}{n_i}\right) - V_{ch}} - \sqrt{2v_T \ln\left(\frac{N_{ch}}{n_i}\right)}\right), \quad (4)$$

where γ is the body-effect parameter and n_i is the intrinsic carrier concentration.

This model is valid for any cross-section shape and depends only on four terms: (i) A_{ch}, denoting the area of cross-section the channel i.e. the area of the blue region in Figure 1(b), (ii) W_{eff}, denoting the effective width of the channel for carrier transport, i.e., the perimeter of the blue region in Figure 1(b), (iii) N_{ch}, representing the doping in the channel, and (iv) C_{ins}, representing the insulator capacitance per unit length, i.e. the capacitance of the yellow region in Figure 1(b), assuming the length (into the paper) is unity. Table 1 provides some examples of calculating these four terms.

Table 1. Model parameter examples. R denotes the radius of cylindrical nanowire. H_{Fin} and T_{Fin} are the height and thickness of fin, and r_c is the radius of curvature of corners in GAAFETs.

	Double-gate	Tri-gate	Cylindrical nanowire	GAAFET/Nanosheet
W_{eff}	$2H_{Fin}$	$T_{Fin}+2H_{Fin}$	$2\pi R$	$2(W_{Si}+T_{Si})+(2\pi-8)r_c$
A_{ch}	$H_{Fin}T_{Fin}$	$H_{Fin}T_{Fin}$	πR^2	$W_{Si}T_{Si}+(\pi-4)r_c^2$
C_{ins}	$2H_{Fin}\frac{\epsilon_{ins}}{T_{ins}}$	$W_{eff}\frac{\epsilon_{ins}}{T_{ins}}$	$2\pi\epsilon_{ins}/\ln\left(1+T_{ins}/R\right)$	$W_{eff}\frac{\epsilon_{ins}}{T_{ins}}$
N_{ch}	Channel doping	Channel doping	Channel doping	Channel doping

In Equation (1), the three terms on the right hand side define the behavior of the charge density in the channel. The linear term dominates in strong inversion, the second term dictates weak-inversion and the third is for the moderate inversion region. This equation, therefore, models the behavior of the channel charge accurately for all bias regions [15]. The core transport equation is the well-known drift-diffusion model [2], given as

$$I_{ds}=-\mu v_T^2\frac{C_{ins}}{L_g}\int_{q_{m,S}}^{q_{m,D}}q_m\frac{dv_{ch}}{dq_m}dq_m$$
$$\Rightarrow I_{ds}=-\mu v_T^2\frac{C_{ins}}{L_g}\left[\frac{q_{m,S}^2-q_{m,D}^2}{2}-2\left(q_{m,S}-q_{m,D}\right)-q_H\ln\left(\frac{q_H-q_{m,S}}{q_H-q_{m,D}}\right)\right] \tag{5}$$

where $q_H=(1/r_N-q_{dep})$. Also $q_{m,S}$ and $q_{m,D}$ are the normalized mobile charge densities at the source and the drain ends, respectively. Second order effects (like various short channel effects) are added on top of Equation (5) [2]. For GAAFET devices with multiple channels per fin, the model scales the calculated quantities appropriately to get the correct terminal characteristics.

3. GAAFET Module

The BSIM-CMG framework has the ability to simulate GAAFETs [2,15]. Recently, however, a few important new code modules have been added to capture the GAAFET specific effects like geometry dependent quantum effects and mobility degradation [13,14]. A new parasitic capacitance network has also been added to capture the effects of the GAAFET structure. In the following subsections we will discuss the most significant GAAFET specific physics that affect the core model behavior.

3.1. Electrostatics

BSIM-CMG, through a geometry module (GEOMOD=5), can calculate accurate values of A_{ch}, C_{ins} and W_{eff}; which are then used in the core model to get the electrostatic behavior, as described in Section 2. The calculation of A_{ch} and W_{eff} include the effects of rounded corners (Figure 1). This model also has the ability to accurately simulate multiple GAA bodies in a single fin (stack). The user can specify various geometry details like the width and thickness of the GAA bodies, the separation between GAA bodies, the number of GAA bodies per fin, fin height etc. The model takes all this geometry information to calculate the electrostatics accordingly. The model can also account for geometry variation among the GAA bodies inside a single fin. In addition to accounting the aforementioned geometry variations, the model further supports Monte Carlo circuit simulation to account for the stochastic device geometric variations that may be encountered in manufacturing.

A significant impact of the confined channel of GAAFETs is the quantum confinement effect on the density of states of silicon. This affects the bias dependence of the channel mobile charge; which in turn affects all device characteristics. To understand the various quantum mechanical effects that play a role, consider the charge in a semiconductor, which can be written as

$$Q=q\sum_i\int_{E_i}^{\infty}g_{D_i}\cdot\frac{1}{1+\exp\left(\frac{E-E_f}{qv_T}\right)}dE$$
$$=\sum_i N_{C_i}F_{\frac{D_i}{2}-1}\left(\frac{E_f-E_i}{qv_T}\right) \tag{6}$$

where g_{Di} is the density of states for the *i*th subband, E_f is the fermi energy and $F_j()$ is the Fermi integral

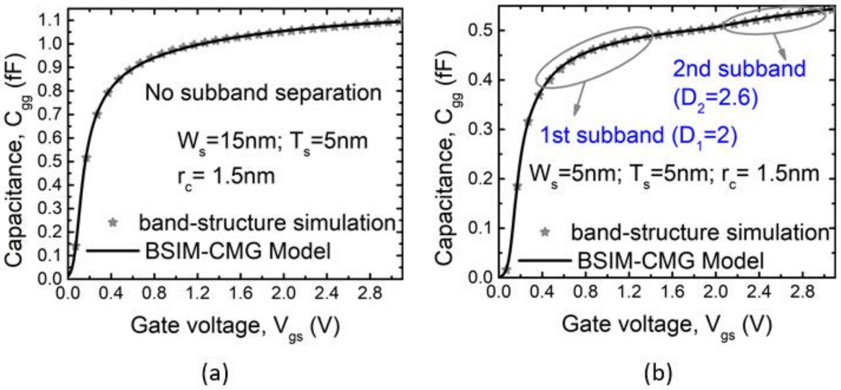


Figure 3. The plots show the gate capacitance with varying gate voltage; for channel thickness of 5nm. For larger cross-sections, like in (a), the confinement is non-existent. Extreme width confinement, as shown in (b), results in a small effect of subband separation.

of order *j*. The term, N_{C_i}, is given as

$$N_{C_i} = q \frac{D_i}{(2\pi)^{\frac{D_i}{2}} \Gamma\left(1+\frac{D_i}{2}\right)} \frac{m_i^{*\frac{D_i}{2}}}{\hbar^{D_i}}, \qquad (7)$$

where D_i is the electrostatic dimension for the i^{th} subband, m_i^* is the effective mass of the i^{th} subband and $\hbar$ is the Planck's constant. In BSIM-CMG, the user is allowed to choose up to 3 subbands and can modify individual subband parameters (refer Table 2).

With changing cross-section, the electrostatic dimension D_i changes. It was recently pointed out that while 1D and 2D are popular and important special cases of quantum confined state, the electrostatic dimension can be a continuous variable. BSIM-CMG is the first compact model that accounts for this fact and can therefore accurately model GAAFET for continuously variable width, W_{Si} [13]. For very confined channels, the system generally has lower dimension. For example, thin and wide channels behave as 2D systems whereas thin and narrow channels are confined in the width direction also, resulting in a 1D behavior. With decreasing confinement, the dimension gradually changes to higher values (2D/3D). This behavior is shown in Figure 3, Figure 4, and Figure 5, where the plots show capacitances (which mimic the density of states) for various cross-sections. As confinement reduces, the dimension shifts from lower to higher values.

Figure 4 and Figure 5 also shows peaks and valleys in the capacitance. These occur due to subband separation. For very confined devices, the conduction band splits up into subbands resulting in peaks in the density of states; which are reflected in the capacitance plots. With increasing confinement, the subband energies increase and they move further apart as illustrated in Figure 5. For larger cross-sections the subband energies reduce and they come closer in energy; forming continuous conduction band. The subband model has been discussed in detail in [13].

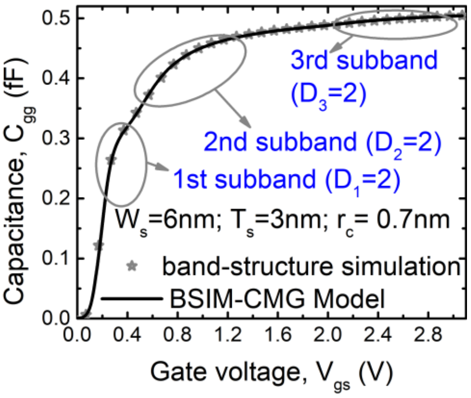


Figure 4. The plot shows the gate capacitance with varying gate voltage; for channel thickness of 3nm. For confined widths, the subband effects are quite pronounced and the overall electrostatic dimension reduces to 2D.

Figure 6 (a) shows the variation of the electrostatic dimension with changing GAAFET width for 2nm thick channels (black line). As confinement reduces with increasing width, the dimension changes from 1D to 2D. The maximum dimension is restricted by the thickness confinement (2nm) and is hence limited to 2D. For thicker GAAFET devices (5nm) the maximum dimension goes up to 3D, as shown by the blue lines. Figure 6 (b) shows the variation of the second subband energy for different cross-sections.

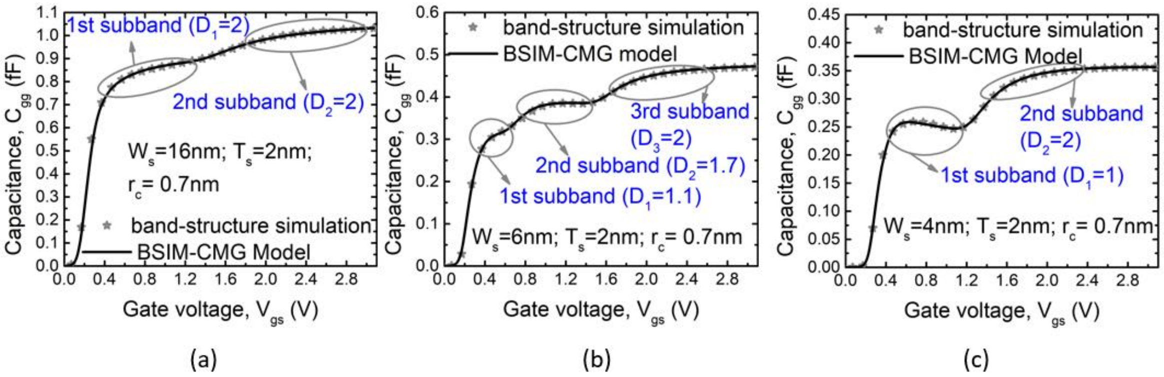


Figure 5. The plots show the gate capacitance with varying gate voltage; for different GAAFET thickness=2nm. The confinement changes from 2D to 1D with decreasing GAAFET width.

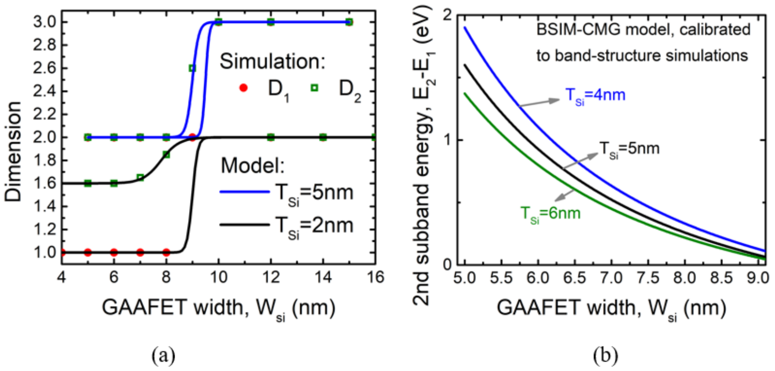


Figure 6. (a) Variation of dimensions for the first and second subbands with thicknesses of 2nm (black lines) and 5nm (blue lines). (b) Variation of the second subband energy, with respect to the first subband energy, for various GAAFET widths and thicknesses.

The capacitance (or charge) also depends on the effective mass, as shown in Equation (6) and Equation (7). The effective mass changes with confinement, and so does the bandgap. The effective mass contributes not only to the charge but also affects mobility. However, the effective mass formulations used in the electrostatics and transport are different. For both these effective mass calculations, we have parameters to modify the geometry dependence based on the device type, material, etc. The geometry dependence of the effective mass for the *i*-th subband in electrostatics calculations is given as

$$m_i^* = m^* + \frac{\Delta_{0,i}}{T_{Si}^{\alpha_m}\left(1+\kappa_m W_{Si}^{\frac{\gamma_0}{T_{Si}^{\beta_m}}}\right)}, \qquad (8)$$

where $\Delta_{0,i}$, γ_0, α_m, β_m and κ_m are device dependent parameters. $\Delta_{0,i}$ can be used as a fitting parameter to tune the variation of effective mass for each subband. Note that the variation of effective mass in Silicon is quite complex since longitudinal and transverse masses react differently to confinement. However, for compact modeling purposes, we use have developed a single expression for geometry dependence of effective mass for electrostatics which has been described in Equation (8) [13].

The bandgap on the other hand plays a role in deciding the threshold voltage. With increasing confinement, both the bandgap and the effective mass increase. This increases the threshold voltage and reduces the amount of charge at a given voltage; as can be seen in Figure 5, Figure 4, and Figure 3.

Another key requirement from the compact model is accuracy for derivatives of charges and currents. The peaks and valleys due to quantum confinement lead to multiple secondary peaks in the derivatives of charges. It is important that the compact model captures this to ensure high accuracy for analog/RF simulations. We have developed and tested our model up to the seventh derivative of charge to ensure high accuracy in non-linearity and harmonics simulations. Figure 7 shows the model results for multiple orders of derivatives along with the simulation results to validate this.

The impact of confinement can also be seen in terminal currents. Figure 8 shows the drain-to-source

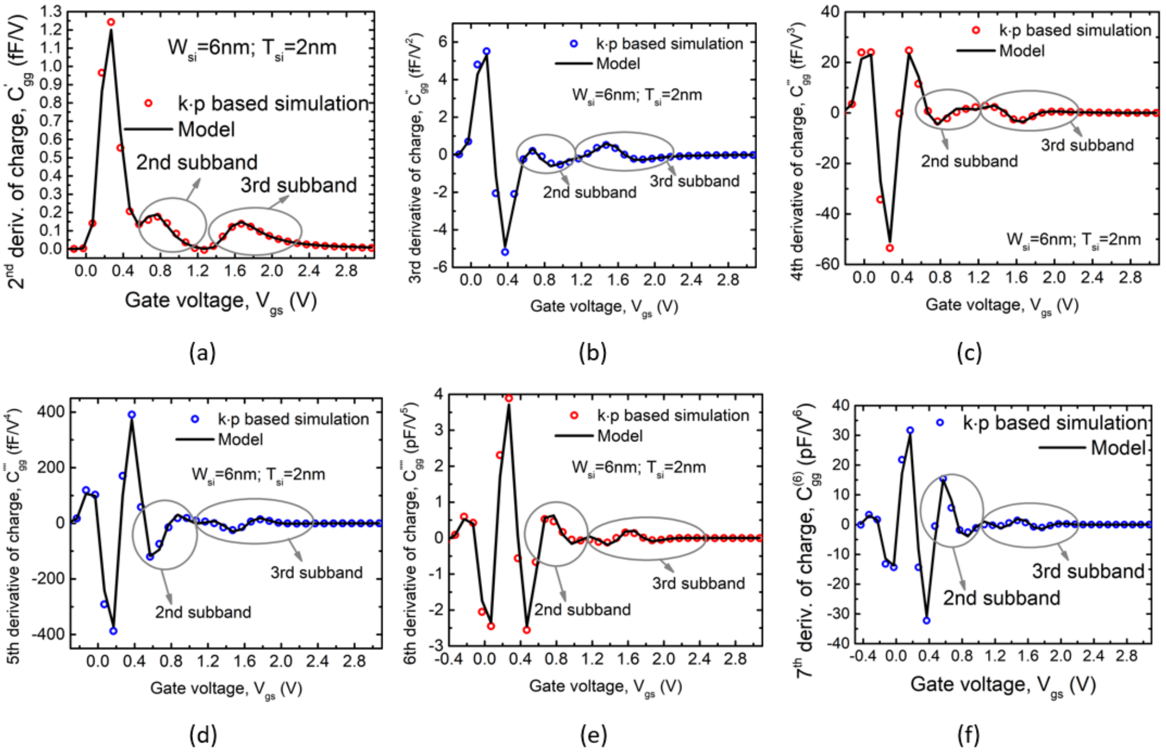


Figure 7. Derivates of charge from 1st to 7th order showing the accuracy of the model for higher order derivatives.

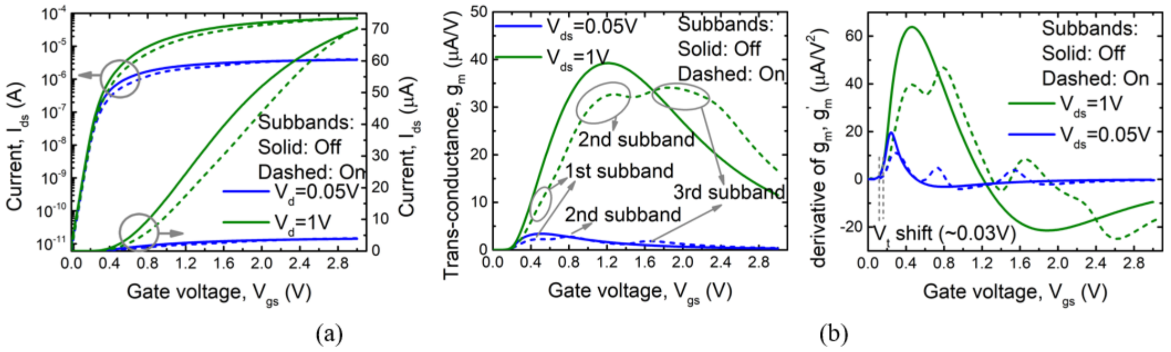


Figure 8. Variation of (a) drain current, (b) transconductance and (c) derivative of transconductance with gate voltage. The solid and dashed lines are the simulation results with and without quantum confinement effects.

current along with the transconductance and the derivative of the transconductance for W_{Si}=6nm and T_{Si}=2nm. The simulation has been done with a constant mobility to remove the effects of confinement on mobility. Impact of electrostatic confinement can be clearly seen in plots. Not only does the current reduce due to lower density of states, the effect of subband separation is also seen as distinct peaks and valleys in the derivatives.

3.2. Transport

In the BSIM-CMG framework, all the transport physics is captured through the concept of effective mobility(μ)[14]. The field dependence of mobility is captured through

$$\mu = \frac{\mu_{0,eff}}{1+\alpha E_{eff}^{\beta}}, \tag{9}$$

where $\mu_{0,eff}$ is the effective mobility at low transverse electric field, α and β are parameters and E_{eff} is the effective transverse electric field. For GAAFETs (as well as FinFETs) the effective mobility is dependent on the Silicon thickness, as shown in Figure 9. Note that the mobility, in general, reduces with increasing confinement. There are multiple factors that contribute to the geometry dependence of mobility, which have been captured through the concept of effective mass. From Figure 9, we can see that high confinement results in the reduction of the effective mobility (μ) which can be modeled by an increasing effective mass. This can be captured through the following equation [14]

$$\mu_{0,eff} = \mu_0 \frac{m_0}{m^*}; \quad \frac{m^*}{m_0} = S_m \frac{\kappa_1 + \sqrt{\kappa_1^2 + 4A\kappa_2}}{2\kappa_2}, \qquad (10)$$

where $\kappa_1 = m_0 E_{g,bulk} T_{Si}^2 - A$ and $\kappa_2 = m_0 (E_{g,bulk} + B) T_{Si}^2$. Here m_0 is the rest mass of an electron, μ_0 is a parameter representing the mobility with effective mass = m_0. $B = 2h^2/(m_0 a_0^2)$ where h is the Planck constant and a_0 is the lattice constant. Also, $E_{g,bulk}$ is the band-gap for bulk Silicon and $A \sim B/64$. S_m is a scaling factor used to tune the dependence for different materials, device types and dopings.

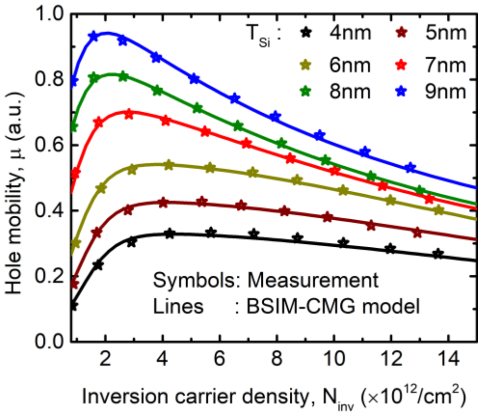


Figure 9. Variation of mobility with inversion carrier density for different GAAFET thicknesses. The measurements are for a p-type device [9]. Mobility reduces with reduction in thickness because of the increase in effective mass with increasing confinement. Moreover, the field dependence of the mobility also changes with reduction of the thickness.

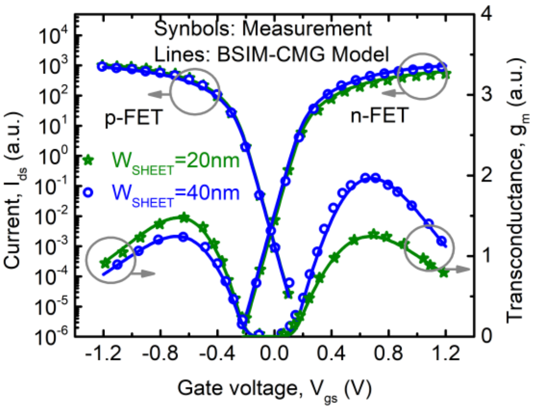


Figure 10. Effect of different mobilities at the sidewall and the top/bottom surface. The mobility scales differently with width scaling for n-type and p-type devices since the ratio of mobilities at the sidewall and top/bottom surface are different for electrons and holes.

The change in effective mass is not enough to capture the geometry dependence of mobility. It is important to note that the field dependence of mobility (high inversion charge) also changes with increasing confinement. This region is dominated by phonon-scattering and surface roughness scattering. This has been captured by including geometry dependence in α and β terms of Equation (9) [14].

Another phenomenon of geometry dependent mobility variation specific to GAAFETs is the effect of the different crystal orientations of the top/bottom surface and the sidewalls. Since these two surfaces are oriented differently, the mobilities for the sidewall and the top/bottom surfaces are different. This leads to the mobility depending on the width as well as thickness of the GAA body and the scaling being a function of the width and thickness. Moreover, the ratio of the mobility of the sidewall to that of the top/bottom surface (η_μ) may be less than unity for electrons and more than unity for holes; leading to completely opposite scaling trends for n-type and p-type devices, as shown in Figure 10. This effect has also been captured in the latest BSIM-CMG GAAFET model [14] as

$$\mu_{eff} = \mu_{top/bottom} \left[\frac{W_{Si}}{W_{Si}+T_{Si}} + \underbrace{\frac{\mu_{sidewall}}{\mu_{top/bottom}}}_{\eta_\mu} \frac{T_{Si}}{W_{Si}+T_{Si}} \right]. \qquad (11)$$

The final expression for mobility is given as [14]

$$\mu = \frac{\mu_0}{1+\alpha\left(W_{Si}, T_{Si}\right) E_{eff}^{\beta(W_{Si}, T_{Si})}} \frac{m_0}{m^*} \left[\frac{W_{Si}}{W_{Si}+T_{Si}} + \eta_\mu \frac{T_{Si}}{W_{Si}+T_{Si}} \right]. \qquad (12)$$

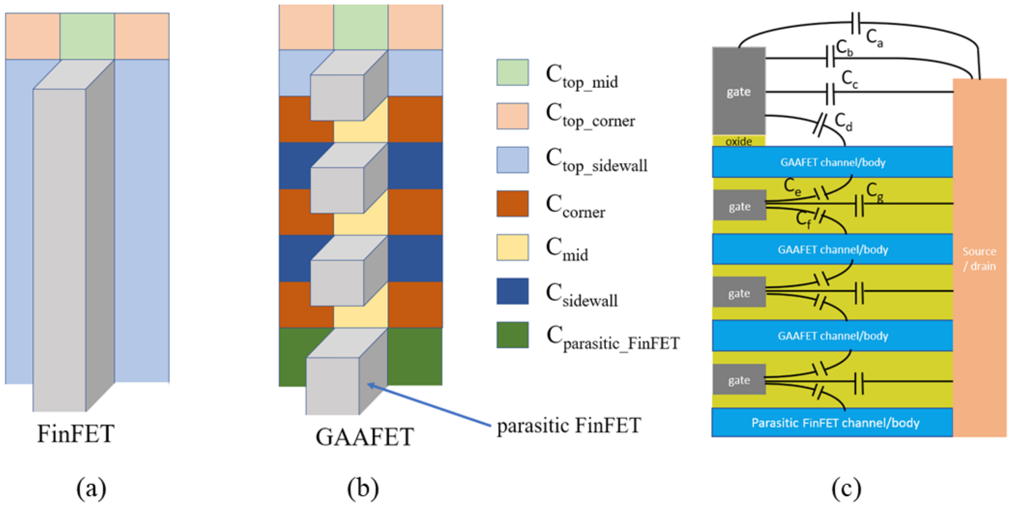


Figure 11. Graphical representation of the parasitic capacitance components for FinFETs and GAAFETs. As can be seen, the GAAFET structure has additional components due to the multiple channels per fin.

4. Parasitic Capacitance

BSIM-CMG includes models for calculation of parasitic capacitances for various device geometries. Accurate modeling of parasitic capacitances plays a crucial role in the accurate analog, digital and RF simulations. Since there are differences between the device structures of FinFETs and GAAFETs, as shown in Figure 1, the latest BSIM-CMG model also has a specific module (CGEOMOD=3) for accurate parasitic capacitance calculation for GAAFET devices. This module takes into account the various structural details of the fin as well as the GAA channels inside it to calculate the various parasitic capacitance elements. The model has the ability to account for multiple GAA channels per fin as well as the parasitic FinFET, indicated in Figure 11(b) which shows the various components of parasitic capacitance for FinFETs and GAAFETs. The GAAFET structure has a more complex parasitic capacitance network because of the multiple GAA channels per fin.

Some of the fringe capacitance components are explicitly shown in Figure 11(c). Due to the curved 3D structure of FinFET and GAAFET channels, the corner components are different from the central ones. Moreover, the GAAFET structure has total six components of fringe capacitance per channel as opposed to only three components for the fin in case of FinFETs. For example, C_d has three components: one for the central region and two for the two corners, as shown in Figure 11(b). This is also true for C_e, C_f, C_g etc. The model for fringe capacitances is derived by summing over the capacitances of small area elements as

$$\Delta C = \epsilon \frac{\Delta A}{d}, \qquad (13)$$

where ΔC is the capacitance corresponding to the infinitesimal area element ΔA, ϵ is the effective permittivity of the insulating material and d is the effective thickness of the insulator. Note that the structure of these devices often does not result in simple parallel-plate capacitance scenarios with straight field-lines. In most cases, the two surfaces of the capacitor are at some angle (mostly orthogonal) to each other and the field-lines curve from one surface to the other. In such cases, the effective d is calculated using the length of the field-line assuming that the field-lines follow an ellipse, as shown in Figure 12 [2]. For orthogonal surfaces, the effective distance is a quarter of the perimeter of an ellipse given by

$$\text{perimeter of ellipse} = 2\pi\sqrt{\frac{a^2+b^2}{2}}, \qquad (14)$$

where a and b are the length of the major and minor axes of the ellipse, respectively.

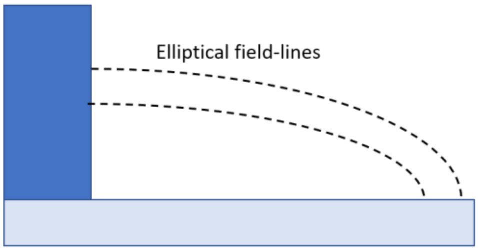


Figure 12. Graphical representation capacitance calculation for orthogonal surfaces.

Table 2. Selected BSIM-CMG parameters used in GAAFET modeling.

Parameters	**Description**
FPITCH	Fin pitch
TMASK	Height of the hard mask on top of the fin
TGATE	Gate height on top of hard mask
HEPI	Height of the raised source/drain on top of the fin
TSILI	Thickness of the silicide on top of the raised source/drain
WGAA	Width of GAA channel (represented by W_{Si} in Figure 1b)
TGAA	Thickness of GAA channel (represented by T_{Si} in Figure 1b)
DWS1/DWS2/DWS3	Rounded corner correction for total channel perimeter of the 1st/2nd/3rd/ GAAFET; in case there are multiple GAAFETs per fin
DACH1/DACH2/DACH3	Rounded corner correction for total channel area of the 1st/2nd/3rd/ GAAFET; in case there are multiple GAAFETs per fin
TSUS	Distance between multiple GAAFETs per fin
NGAA	Number of GAA per fin
HPFF	Height of parasitic FinFET
U0ETAWSC	Ratio of the mobility of the sidewall to that of the top/bottom surface
EGBULK	Bulk band-gap

The calculation of overlap capacitances also changes from FinFET to GAAFET since the overlap length changes because of the GAAFET structure. Also, multiple GAAFET channels in a single fin requires the overlap capacitance of a single GAAFET be scaled by the total number of channels per fin to ensure that the terminal characteristics are captured correctly [2].

5. Model Parameters

BSIM-CMG provides the model user with carefully implemented model parameters that can be adjusted using software tools to match the model to manufactured transistors very accurately. This crucial step is performed by the foundry of the fab of an integrated device manufacturer. One may say that a device model is the model code, such as BSIM-CMG, plus a specific parameter value set. For example, the difference between Samsung 3nm GAA transistors and TSMC 2nm GAA transistors are captured and represented by two difference sets of the BSIM-CMG parameters. These parameters, about forty in number for the GAA related effects, in conjunction with the device information that the IC designer specifies, such as the width of the GAA channel and the length of the GAA gate, are used by computer-aided IC design tools to simulate, design and optimize circuits. Some of the key parameters for GAAFET devices are specified in Table 2.

Simulation speed is also a key characteristic of a good compact models. BSIM-CMG includes all of necessary physics while rapidly calculating all the terminal currents and charge (for capacitive currents) for any given terminal voltages. The speed allows the use of Monte Carlo circuit simulation to account for random device variations encountered in manufacturing. The compact model also provides some parameters to allow the model user to optimize their simulation accuracy versus time to best suit their requirements.

6. Conclusion

We have presented the BSIM-CMG compact model framework; with emphasis on the modeling of GAAFETs. This compact model has been extensively used by the semiconductor industry for FinFET based IC designs. We have discussed the model core which forms the backbone for all the calculation. We have also discussed the latest modules that capture the potentially strong effects of quantum confinement on silicon density of states and transport in GAAFET devices. This model is the industry standard compact model for simulating and designing GAA ICs, libraries and IPs.

References

[1] X. Huang, W-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y-K. Choi, K. Asano, V. Subramanian, T-J. King, J. Bokor, C. Hu, "Sub 50-nm FinFET: PMOS," IEDM Technical Digest, Washington, DC, pp. 67-70, December 5-8, 1999.

[2] Y. S. Chauhan et al., "FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard".

New York, NY, USA: Academic, 2015, doi: 10.1016/B978-0-12-420031-9.09994-2.

[3] H. Mertens et al., "Vertically stacked gate-all-around Si nanowire CMOS transistors with dual work function metal gates," in IEDM Tech. Dig., Dec. 2016, pp. 19.7.1–19.7.4, doi: 10.1109/IEDM.2016.7838456.

[4] M. Karner et al., "Vertically stacked nanowire MOSFETs for sub-10 nm nodes: Advanced topography, device, variability, and reliability simulations," in IEDM Tech. Dig., Dec. 2016, pp. 30.7.1–30.7.4, doi: 10.1109/IEDM.2016.7838516.

[5] Y. Jiang et al., "Performance breakthrough in 8 nm gate length Gate-AllAround nanowire transistors using metallic nanowire contacts," in Proc. Symp. VLSI Technol., Jun. 2008, pp. 34–38, doi: 10.1109/VLSIT.2008.4588553.

[6] Y. Cui et.al., "High performance silicon nanowire field effect transistors," Nano Lett., vol. 3, no. 2, p. 149–152, 2003, doi: 10.1021/nl025875l.

[7] N. Loubet et al., "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET," in Proc. Symp. VLSI Technol., Jun. 2017, pp. T230–T231, doi: 10.23919/VLSIT.2017.7998183.

[8] K. H. Yeo et al., "Gate-all-around (GAA) twin silicon nanowire MOSFET (TSNWFET) with 15 nm length gate and 4 nm radius nanowires," in IEDM Tech. Dig., Dec. 2006, pp. 1–4, doi: 10.1109/IEDM.2006.346838.

[9] C. W. Yeung et al., "Channel geometry impact and narrow sheet effect of stacked nanosheet," in IEDM Tech. Dig., Dec. 2018, pp. 28.6.1–28.6.4, doi: 10.1109/IEDM.2018.8614608.

[10] G. Bae et al., "3 nm GAA technology featuring multi-bridge-channel FET for low power and high performance applications," in IEDM Tech. Dig., 2018, pp. 28.7.1–28.7.4, doi: 10.1109/IEDM.2018.8614629.

[11] A. Dasgupta, A. Agarwal, and Y. S. Chauhan, "Unified compact model for nanowire transistors including quantum effects and quasi-ballistic transport," IEEE Trans. Electron Devices, vol. 64, no. 4, pp. 1837–1845, Apr. 2017, doi: 10.1109/TED.2017.2672207.

[12] A. Dasgupta et.al., "Compact modeling of cross-sectional scaling in gate-all-around FETs: 3-D to 1-D transition," IEEE Trans. Electron Devices, vol. 65, no. 3, pp. 1094–1100, Mar. 2018, doi: 10.1109/TED.2018.2797687.

[13] A. Dasgupta et.al., "BSIM Compact Model for Quantum Confinement in Advanced Nanosheet FETs", IEEE Transactions on Electron Devices, vol. 67, no. 2, 2020.

[14] A. Dasgupta et.al., "Compact Model for Geometry Dependent Mobility in Nanosheet FETs", IEEE Electron Device Letters, vol. 41, no. 3, 2020.

[15] J. P. Duarte et.al., "BSIM-CMG: Standard FinFET Compact Model for Advanced Circuit Design", IEEE European Solid-State Circuit Conference (ESSCIRC), Graz, Austria, Sept. 2015.

[16] J. Wang et.al., "Bandstructure and orientation effects in ballistic Si and Ge nanowire FETs," in IEDM Tech. Dig., Dec. 2005, p. 533, doi: 10.1109/IEDM.2005.1609399.

[17] Y. S. Chauhan et al., FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard. New York, NY, USA: Academic, 2015, doi: 10.1016/B978-0-12-420031-9.09994-2.

[18] B. Sorée, W. Magnus, and G. Pourtois, "Analytical and self-consistent quantum mechanical model for a surrounding gate MOS nanowire operated in JFET mode," J. Comput. Electron., vol. 7, no. 3, pp. 380–383, 2008, doi: 10.1007/s10825-008-0217-3.

[19] S. Venugopalan et.al., "Phenomenological compact model for QM charge centroid in multigate FETs," IEEE Trans. Electron Devices, vol. 60, no. 4, pp. 1480–1484, Apr. 2013, doi: 10.1109/TED.2013.2245419.

Photography & Biography

Avirup Dasgupta is a postdoctoral scholar at the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA. He is the manager of the Berkeley Device Modeling Center and a developer in the BSIM group.

Chenming Hu is currently a Distinguished Professor Emeritus with the University of California at Berkeley, Berkeley, CA, USA. He is also a Board Director of SanDisk Inc., Milpitas, CA, USA, and the Friends of Children with Special Needs, Fremont, CA, USA

J. Microelectron. Manuf. **3**, 20030403 (2020)
doi: 10.33079/jomm.20030403

On the History of the Numerical Methods Solving the Drift Diffusion Model

Bernd Meinerzhagen [*]

Technical University Braunschweig, Germany

Abstract: In 1964 Hermann Gummel published the first numerical solution method for the one-dimensional Drift Diffusion model. In his seminal paper [1] already the nonlinear iteration method and the basics of the discretization method named after him are outlined. Soon after this paper appeared many research groups worldwide tried to solve the Drift Diffusion equations in two and more dimensions applying predominantly general finite element discretization methods which were very popular at these days. Due to this a large variety of different codes solving the multidimensional Drift Diffusion equations based on many different space discretization schemes existed in the seventies. However already in the nineties all Drift Diffusion simulators being of importance for semiconductor device design in industry and academia still used Gummel's nonlinear iteration method but were entirely based on just one specialized space discretization method, which incorporates the basic ideas of the Scharfetter-Gummel discretization scheme [2]. All other codes which were not based on this special space discretization method had nearly vanished already in the nineties and this is still the case today. This paper tries to shed some light on the hidden reasons for this astonishing development.

1. Introduction

In the following the development of the numerical solution methods for the Drift Diffusion (DD) equations will be reviewed, since much can be learned from this development for comparable tasks in the future. Aspects covered well in the literature will be only shortly mentioned by citing the appropriate references. Other aspects that are very important as well but rarely mentioned in literature or even nearly forgotten today will be highlighted. Especially the aspect of preserving the inherent "stability" of the underlying differential equations in terms of monotonicity properties during the discretization and solution process will be carefully examined. The general flow of arguments presented here follows closely those outlined in Reference [3]. Several details have been previously published as well in References [4, 5]. Some hints concerning the Hydrodynamic (HD) model [6] will be given at the end.

2. History of the Numerical Models

Since the quasi Fermi potentials Φ_n and Φ_p (imrefs) for electrons or holes, respectively, cover a much smaller range of numerical values compared to the electron or hole densities n, p, intermediate and final solutions of the drift diffusion system of equations are typically saved by storing the electrostatic potential Ψ, and the imrefs Φ_n and Φ_p instead of Ψ, n and p. Moreover, formulating the DD equations in Ψ, Φ_n, and Φ_p makes it much easier to introduce the solution algorithms that are typically applied. The stationary drift diffusion equations for a homogeneous semiconductor (e.g., silicon) device formulated in these variables have the following form (see References [7, 4, 8, 9] for details and a derivation),

$$\begin{aligned} TP\left(\Psi,\Phi_p,\Phi_n\right) &:= -\nabla_{\boldsymbol{r}}\cdot\left[\kappa\nabla_{\boldsymbol{r}}\Psi\right] \\ &+e\left(n_i\left(\exp\left(\frac{\Psi-\Phi_n}{V_T}\right)-\exp\left(\frac{\Phi_p-\Psi}{V_T}\right)\right)-N_D+N_A\right)=0 \end{aligned} \tag{1}$$

$$\begin{aligned} &TH\left(\Psi,\Phi_p,\Phi_n\right):= \\ &-\nabla_{\boldsymbol{r}}\cdot\left(n_i\exp\left(\frac{\Phi_p-\Psi}{V_T}\right)\mu_p\nabla_{\boldsymbol{r}}\Phi_p\right)-G= \\ &-\nabla_{\boldsymbol{r}}\cdot\left(n_i\exp\left(\frac{-\Psi}{V_T}\right)V_T\mu_p\nabla_{\boldsymbol{r}}\exp\left(\frac{\Phi_p}{V_T}\right)\right)-G \\ &=:\hat{TH}\left(\Psi,\exp\left(\frac{\Phi_p}{V_T}\right),\Phi_n\right)=0 \end{aligned} \tag{2}$$

[*] Address all correspondence to Bernd Meinerzhagen, E-mail: b.meinerzhagen@tu-bs.de

$$\begin{aligned} &TE\left(\Psi,\Phi_p,\Phi_n\right) := \\ &-\nabla_r \cdot \left(n_i \exp\left(\frac{\Psi-\Phi_n}{V_T}\right)\mu_n \nabla_r \Phi_n \right) + G = \\ &+\nabla_r \cdot \left(n_i \exp\left(\frac{\Psi}{V_T}\right) V_T \mu_n \nabla_r \exp\left(\frac{-\Phi_n}{V_T}\right)\right) + G \\ &=: -\hat{TE}\left(\Psi,\Phi_p,\exp\left(\frac{-\Phi_n}{V_T}\right)\right) = 0 \end{aligned} \tag{3}$$

where e is the elementary charge and κ is the permittivity of the different materials for which Poisson's Equation (1) is solved. Moreover n_i is the intrinsic density, N_D, N_A are the ionized donor and acceptor concentrations and μ_n, μ_p the electron and hole mobilities of the semiconductor material with homogeneous band gap within which the electron and hole continuity Equations (3, 2) are solved. In addition, V_T is the thermal voltage and G the generation density within the semiconductor. For simplicity it is assumed that at all contacts the Dirichlet boundary conditions of the ideal Ohmic contact model (see References [4, 9] for details) are valid for all three potentials Ψ, Φ_n and Φ_p and that at all other boundaries homogeneous Neuman type boundary conditions can be applied.

It can be shown that the above system of differential equations has a unique solution provided a number of reasonable assumptions is fulfilled especially for the generation term G. Please refer to Reference [10] for a general theory and to References [11, 12] for special results concerning the DD set of equations.

There is an important property, that deserves special attention.

I. All above operators TP - TE are in divergence form and result in important conservation laws if integrated over a finite volume and after the application of the divergence theorem.

For example integrating *TH*+*TE* over the simulation domain results in Kirchhoff's law for the stationary terminal currents of the device under consideration. If a numerical device model is used inside a circuit simulator it is absolutely mandatory that this law is exactly reproduced by the numerical model. Therefore, it is very important to maintain the validity of such conservation laws in some sense during the discretization process.

The most important solution algorithm for the DD system is an iterative method often addressed as Gummel's nonlinear relaxation method [1]. Assuming the result Ψ_k, $\Phi_{p,k}$, $\Phi_{n,k}$, after iteration k as known, this method evaluates the new approximate solution after iteration k+1 by solving the three boundary value problems (1) - (3) successively as follows:

$$\begin{aligned} TP\left(\Psi_{k+1},\Phi_{p,k},\Phi_{n,k}\right) &= 0\ , \\ TH\left(\Psi_{k+1},\Phi_{p,k+1},\Phi_{n,k}\right) &= 0\ , \\ TE\left(\Psi_{k+1},\Phi_{p,k+1},\Phi_{n,k+1}\right) &= 0 \end{aligned} \tag{4}$$

For the above partiell differential equations it is always assumed that the variable with the highest number of derivatives in the equation is updated and the other variables are kept unchanged. The individual nonlinear equations in (4) are typically solved by Newton's method, which converges very fast and robust even for bad initial solutions, if the underlying equation is nearly linear. Therefore, Newton's method is typically applied for the operators $\hat{TH}$ and $\hat{TE}$ defined in Equation (2) and Equation (3) and not for the operators *TH* and *TE*, since $\hat{TH}$ and $\hat{TE}$ are nearly linear in the new variables

$$\zeta_p := \exp\left(\frac{\Phi_p}{V_T}\right),\ \zeta_n := \exp\left(\frac{-\Phi_n}{V_T}\right)\ , \tag{5}$$

provided carrier generation G has no dominant influence. These new variables are often addressed as Slotboom variables because they were first introduced in Reference [13] and the advantages they have for generating "stable" discretization schemes were probably mentioned in Reference [7] for the first time. Nevertheless it is still possible to calculate the Newton updates based on the numerically more convenient Jacobians of *TH* and *TE* and the variables Φ_p and Φ_n. The only modification necessary for performing the Newton iterations for the more linear operators is to modify the Newton update itself as shown below for the hole continuity equation and the solution function before ($\Phi_{p,b}$) and after ($\Phi_{p,a}$) one Newton step.

$$\begin{aligned} &\Phi_{p,a} = \Phi_{p,b} + \delta\Phi_p \rightarrow \\ &\qquad \Phi_{p,a} = \Phi_{p,b} + V_T \ln\left(1+\frac{\delta\Phi_p}{V_T}\right) \end{aligned} \tag{6}$$

$\delta\Phi_p$ is the Newton update calculated using the Jacobian of *TH* for the variable ϕ_p . For a general report on using alternative solution variables for enhancing convergence please refer to Reference [14].

In order to understand which criteria in addition to the mandatory consistency criterion should be considered for the discretization of the continuous operators *TP* - *TE* it is very useful to look at the Jacobians of *TP*, $\hat{TH}$ and $\hat{TE}$ that are necessary for performing Gummel's nonlinear relaxation method based on Newton's method. These Jacobians are typically evaluated for some existing intermediate solution Ψ_b, $\Phi_{p,b}$, $\Phi_{n,b}$ and operate on the functions $\delta\Psi$, $\delta\zeta_p$, $\delta\zeta_n$. If in addition for the calculation of the Jacobians the dependence of the mobilities on the solution variables is neglected, these Jacobians have the following form.:

$$\begin{aligned}&\frac{\partial}{\partial\Psi}TP\left(\Psi_b,\Phi_{p,b},\Phi_{n,b}\right)\left[\delta\Psi\right]:=-\nabla_{\mathbf{r}}\cdot\left[\kappa\nabla_{\mathbf{r}}\delta\Psi\right]\\&+e\left(n_i\left(\exp\left(\frac{\Psi_b-\Phi_{n,b}}{V_T}\right)+\exp\left(\frac{\Phi_{p,b}-\Psi_b}{V_T}\right)\right)\right)\delta\Psi\end{aligned}\tag{7}$$

$$\begin{aligned}&\frac{\partial}{\partial\zeta_p}\hat{TH}\left(\Psi_b,\zeta_{p,b},\Phi_{n,b}\right)\left[\delta\zeta_p\right]:=\\&-\nabla_{\mathbf{r}}\cdot\left(n_i\exp\left(\frac{-\Psi_b}{V_T}\right)V_T\mu_{p,b}\nabla_{\mathbf{r}}\delta\zeta_p\right)-\frac{\partial}{\partial\zeta_p}G\left(\Psi_b,\zeta_{p,b},\Phi_{n,b}\right)\delta\zeta_p\end{aligned}\tag{8}$$

$$\begin{aligned}&\frac{\partial}{\partial\zeta_n}\hat{TE}\left(\Psi_b,\Phi_{p,b},\zeta_{p,n}\right)\left[\delta\zeta_n\right]:=\\&-\nabla_{\mathbf{r}}\cdot\left(n_i\exp\left(\frac{\Psi_b}{V_T}\right)V_T\mu_{n,b}\nabla_{\mathbf{r}}\delta\zeta_n\right)-\frac{\partial}{\partial\zeta_n}G\left(\Psi_b,\zeta_{p,b},\zeta_{n,b}\right)\delta\zeta_n\end{aligned}\tag{9}$$

All above partial derivatives with respect to solution variables are assumed to be Frechet derivatives on suitable function spaces [15]. If only direct recombination and Shockley-Read-Hall (SRH) recombination [9] is considered for the Frechet derivative of the carrier generation G

$$\begin{aligned}&-\frac{\partial}{\partial\zeta_p}G\left(\Psi_b,\zeta_{p,b},\zeta_{n,b}\right)>0\\&-\frac{\partial}{\partial\zeta_n}G\left(\Psi_b,\zeta_{p,b},\zeta_{n,b}\right)>0\end{aligned}\tag{10}$$

holds. With this additional condition all Jacobians (7)-(9) have important properties that again deserve special attention. They are

II. self adjoint,
III. positive definite,
IV. of monotone typ,

if appropriate boundary conditions are assumed for the function spaces considered [15]. Especially property IV is very important, since it means that for all these Jacobians monotonicity theorems hold (Reference [15], Chapter 23.5) which restrict for example the possible form of the update functions $\delta\Psi$, $\delta\zeta_p$, $\delta\zeta_n$ during the Newton iterations required for Gummel's nonlinear relaxation method very much and enhance the robustness and convergence properties of this solution method decisively.

The mathematical properties of the model after discretization should be as similar as possible to the properties of the continuous model! If this is fulfilled the discrete model is an analogon of the continuous model, even if only coarse grids can be afforded, which is typically the case. This property is very important for the discretization error control on coarse grids. Therefore discretization methods are preferred which are able to conserve conditions I -IV in some discrete sense.

The first 2D DD simulations used a rectangular solution domain and tensor product grids [16, 7] so that standard finite difference discretization methods for tensor product grids could be applied, that conserved most of the conditions mentioned above. However many device cross sections were not rectangular, so that at least at the beginning of multi-dimensional numerical semiconductor device modeling many groups developed discretization methods (see References [17, 18, 19, 20, 5, 21, 22] and citations therein) for non rectangular solution domains based on the finite element approach [23], but it turned out that general finite element methods typically have problems to conserve conditions I-IV [24, 18, 5, 25]. Finally, even for solution domains with complicated polygonal boundaries which cannot be discretized efficiently by a tensor product grid, the method of choice used today by the vast majority of DD and HD simulators is a straight forward generalization of the integration method published in Reference [26], Chapter 6 for 2D tensor product grids. This method is able to conserve conditions I-IV as will be shown below. This generalization for 2D problems was already briefly mentioned in Reference [26] using the early work of Reference [27] as guideline. This general method is addressed as box integration method in Reference [4], as box method in Reference [5] and today often named finite volume method [28]. An especially well documented example of this historical development from the application of general finite element methods to the final exclusive application of the finite volume method is the development of numerical device modeling codes at IBM research. There two groups independently developed two general numerical device modeling codes. One group with a clear focus on the application of general finite element methods [24, 19, 20] and the other group shifting more and more from hybrid finite element/ finite volume discretization methods to the exclusive application of the finite volume method [29, 18]. Both codes were developed over a decade until the beginning of the eighties. Ten years later the finite element code development is not mentioned at all any more in a comprehensive review paper about the TCAD development at IBM with more than 200 citations [30].

Possibly the first mathematical analysis of the general box integration method for the DD model in 3D is due to Reference [31]. The box integration method can be interpreted as a finite element method [5], but the grid elements for this method (e.g. triangles) cannot be considered as the basis from which the discretization method proceeds like it is the case for a general finite element method [23] but instead the elements are constructed in a unique way based on the predefined grid points. The method is based on the construction of two dual grids the Voronoi diagram and the Delaunay tessellation. The early work of the two Russian mathematicians M.G. Voronoi [32] and B.N. Delone [33] is typically cited in this context but the method in 2D is even much older and has been rediscovered various times. Figure 1 below is used to explain the basic principles of this method. The n grid points P_k (n = 9 in this example) with the coordinates $\boldsymbol{r}_k$ are considered as given. In a first step for each point with index k its Voronoi volume V_k is constructed as the set of all points in space that are closer to P_k than to any other grid point.

$$V_k := \left\{ \boldsymbol{r} \mid \|\boldsymbol{r}-\boldsymbol{r}_k\| \le \|\boldsymbol{r}-\boldsymbol{r}_j\|,\ j=1,..,n,\ j\neq k \right\} \tag{11}$$

Moreover for each grid point P_k its environment S_k is defined by

$$S_k := \left\{ j \mid M\left(V_k \cap V_j\right) \neq 0,\ j=1,..,n,\ j\neq k \right\}. \tag{12}$$

M indicates the 1D Lebesgue measure M_1 for 2D problems and the 2D Lebesgue measure M_2 for 3D problems. With this environment the boundary of V_k is given by

$$\delta V_k := \bigcup_{j\in S_k} (V_k \bigcap V_j) \tag{13}$$

The union of all Voronoi volume boundaries defines the Voronoi diagram. Based on this diagram the dual grid structure of the Delaunay grid is defined by defining the set of edges D for this grid by the straight lines $l_{i,k}$ between the points P_i and P_k for which $M(V_k \bigcap V_i) \neq 0$ and i and k vary between 1 and n with $i \neq k$. Please note that the $l_{i,k}$ are oriented curves starting at P_i and ending at P_k. If only the edges of the Delaunay grid without orientation are important $l_{i,k}$ and $l_{k,i}$ can be considered as equal. The typical elements in such a Delaunay grid for 2D problems are triangles and tetrahedra for 3D problems. But 2D and 3D rectangular tensor product grids are special cases for this method and fit perfectly into this framework. The DD and HD problems are typically formulated as boundary value problems on some finite region Ω and in order to incorporate Newman type boundary conditions into the framework of box integration in a natural manner the boundary $\delta\Omega$ is typically assumed to be composed of edges (faces in 3D) of the delaunay grid elements. In the given example in Figure 1, $\delta\Omega$ is the closed polygonal line composed of $l_{1,3}$, $l_{3,5}$, $l_{5,8}$, $l_{8,9}$, $l_{9,7}$, $l_{7,2}$ and $l_{2,1}$ and Ω is the interior of this closed polygonal line. The Voronoi volumes for the grid points on $\delta\Omega$ are typically unbounded but for box integration bounded boxes are mandatory. This leads to the following definitions:

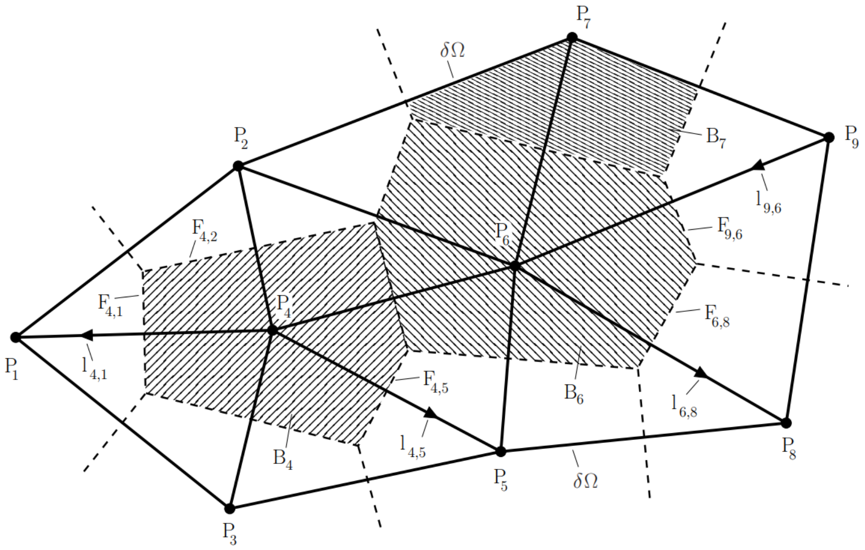


Figure 1. Voronoi diagram (dashed lines) and Delaunay tessellation (thick drawn lines) for a 2D example with 9 grid points.

$$B_k = V_k \bigcap (\Omega \mathrm{U} \delta\Omega) \qquad (14)$$

$$F_{k,j} = \delta V_k \bigcap \delta V_j \bigcap \Omega \qquad (15)$$

Clearly $M(F_{k,j}) \geq 0$ and $F_{k,j}$ is only relevant if $l_{k,i}$ is part of the Delaunay grid. If no additional constraints are fulfilled it can happen that $M(F_{k,j}) = 0$ for some edge of the Delaunay grid. This is typically not good for the consistency of the discretization method and mostly avoided by constructing the grid in such a manner that for example for 2D problems like the given example the Delaunay triangles that have some common boundary edge with $\delta\Omega$ have all interior angles smaller than 90 degrees. Such triangles are called acute or nonobtuse. See for example Reference [34] for algorithms generating grids in such a manner. If this additional condition is fulfilled $M(F_{k,j}) > 0$ for all edges of the Delaunay grid holds and only the boxes B_k of grid points on the boundary $\delta\Omega$ share some boundary with $\delta\Omega$ like B_7 for the given example. Similar additional conditions with comparable consequences are considered as well for 3D problems [31]. For the box integration process presented here it is not necessary that all triangles are nonobtuse for preserving condition IV but for finite element discretization schemes this condition must be fulfilled [34, 5].

The boundary value problems TP, $\hat{T}H$ and $\hat{T}E$ defined in Equations (1) - (3) and considered as individual problems that are solved separately have the following common form:

$$\begin{aligned} &-\nabla_{\boldsymbol{r}} \cdot (a(\boldsymbol{r},\nabla_r u,u)\nabla_r u)+ \\ &f(\boldsymbol{r},\nabla_r u,u)=0 \end{aligned} \qquad (16)$$

$u(\boldsymbol{r})$ is the solution variable. Therefore u is either Ψ or ζ_p or ζ_n. Morover $a(\boldsymbol{r},\nabla_r u,u) > 0$ holds always. The dependence of a and f on $\nabla_r u$ and u considers the typical physical models for the mobilities and the generation rate and their dependence on the solution variables [9]. The first step always performed for the box method is to integrate Equation (16) over the box B_k for each grid point P_k, which is not determined by a Dirichlet boundary condition. Moreover the divergence theorem is used to transform the suitable parts of the integral over B_k into integrals over δB_k. This yields:

$$\begin{aligned} &-\sum_{j \in S_k} \int_{F_{k,j}} a(\boldsymbol{r},\nabla_r u,u)(\nabla_r u \cdot \boldsymbol{v}_{k,j})dF \\ &+\int_{B_k} f(\boldsymbol{r},\nabla_r u,u)dV = 0 \end{aligned} \qquad (17)$$

$\boldsymbol{v}_{k,j}$ is a unit vector having the same orientation as $l_{k,j}$. The above formulation is for the 3D case, where the first integrals integrate fluxes over an area and the remaining part is a volume integral. In the 2D case the first integrals are line integrals and the second part is an area integral. If $M(B_k \bigcap \delta\Omega) \neq 0$ it can typically be assumed that for this part of the boundary $(B_k \bigcap \delta\Omega)$ a Newman type boundary condition holds such that the flux through this part of the boundary is zero. This applies for instance for B_7 in the example, if for P_7 no Dirichlet boundary condition is given. It is clear that

$$\begin{aligned}&\int_{F_{k,j}} a(\boldsymbol{r},\nabla_{\boldsymbol{r}}u,u)(\nabla_{\boldsymbol{r}}u\cdot\boldsymbol{v}_{k,j})dF\\ &-\int_{F_{j,k}} a(\boldsymbol{r},\nabla_{\boldsymbol{r}}u,u)(\nabla_{\boldsymbol{r}}u\cdot\boldsymbol{v}_{j,k})dF\end{aligned} \qquad (18)$$

where the first integral is related to P_k and the box integration over B_k, whereas the second integral is related to the point P_j and the box integration over B_j. During the discretization process the integrals in Equation (17) are typically approximated independently by difference approximations and quadrature rules such that a consistent discrete approximation is generated. For conserving property **I** it is important that Equation (18) holds exactly after discretization. If this is the case the discretized formulas for two points P_j and P_k with $k \in S_j$ (like P_4 and P_6 in the example) can be summed leading to an expression where the sum of the discretized integrals over B_k and B_j is represented by the discretized flux integrals over the boundary of $B_k \bigcup B_j$ which is $((\bigcup_{i\in S_k} F_{k,i})\bigcup(\bigcup_{i\in S_j} F_{j,i})\bigcup((B_k \bigcup B_j)\bigcap \delta\Omega))\setminus F_{k,j}$. The resulting expression can be interpreted as a discrete version of the divergence theorem for $B_k \bigcup B_j$ and its boundary. If the discretized problem is solved exactly this discretized version of the integral theorem holds exactly and does not depend on any discretization error. Relations of this kind are very helpful for checking the global numerical accuracy and the consistent calculation of for instance terminal currents.

In order to study how to preserve properties **II-IV** during the discretization process equation (16) is simplified again by neglecting the dependence of a on the solution variable u and considering only the dependence of direct and SRH generation on the solution variable. Thus Equation (16) simplifies to

$$-\nabla_{\boldsymbol{r}}\cdot(a(\boldsymbol{r})\nabla_{\boldsymbol{r}}u)+f(\boldsymbol{r},u)=0 \qquad (19)$$

and $\frac{\partial}{\partial u}f(\boldsymbol{r},u)>0$ holds always. If the discretization preserves property **I** the discretized integral over $F_{k,j}$ in Equation (17) can be considered as a function $G(l_{k,j})$. This implies that G depends as well on everything clearly connected to $l_{k,j}$ like P_k and P_j . Since Equation (18) shall be preserved $G(l_{k,j})=-G(-l_{k,j})=-G(l_{j,k})$ must hold. Lets assume that the discrete solution is represented by a vector $\boldsymbol{u}$ with n entries u_j and each u_j is the discrete approximation of the function u at the point P_j. The Jacobian of the equation system after discretization should be self adjoint (symmetric). This requires the condition $\frac{\partial}{\partial u_j}G(l_{k,j})=\frac{\partial}{\partial u_k}G(l_{j,k})$. There are not too many alternatives left if these above two conditions must be fulfilled simultaneously. One discretization formula, for which both conditions hold, is

$$G(l_{k,j})=-\frac{M(F_{k,j})}{M_1(l_{k,j})}\bar{a}(l_{k,j})(u_j-u_k). \qquad (20)$$

Here $\bar{a}(l_{k,j})$ is a suitable mean value of $a(\boldsymbol{r})$ that can be considered as a function of $l_{k,j}$ and for which $\bar{a}(l_{k,j})=\bar{a}(l_{j,k})>0$ must be satisfied. If in addition the integral over the box B_k in Equation (17) is discretized using the simplest quadrature formula

$$\int_{B_k} f(\boldsymbol{r},u)dV \approx \nu(B_k)\,f(\boldsymbol{r}_k,u_k), \qquad (21)$$

the strict diagonal dominance of the Jacobian matrix of the discrete system is guaranteed as well. ν indicates the 2D Lebesgue measure for 2D problems and the 3D Lebesgue measure for 3D problems. So far only the discretization for all points which are not given by a Dirichlet boundary condition has been studied. The set of indices of these points should be given by S_B. Moreover S_D contains all indices of points that are given by a Dirichlet condition. The Jacobian entries for the latter points are simply 1 for the main diagonal and 0 for the other entries. In summery the Jacobian of the discretized boundary value problem (19) is strictly diagonal dominant, all main diagonal elements are strictly positive and all other elements are negative or zero. Such matrices are positive definit and M-matrices as well, which means that their inverse matrix has only elements that are positive or zero [26]. These properties are very beneficial for a large number of solution algorithms solving linear equations involving matrices. The convergence of iterative methods like the Jacobi or Gauss-Seidel methods is guaranteed [26], semi-iterative methods like conjugated gradient algorithms work well [35] and even Gaussian elimination profits because a pivot element search is not necessary and the accumulation of the rounding error during elimination is well controlled. Finally and probably most important the discrete system introduced above is of monotone type, which means that important stability inequalities can even be derived for the maximum norm, which is the most important norm for practical applications. For example in Reference [36], Chapter 4.2.2, the following is proven. If for all $k \in S_B$

$$\sum_{j\in S_k} -\frac{M(F_{k,j})}{M_1(l_{k,j})}\bar{a}(l_{k,j})(\delta u_j - \delta u_k) + \nu(B_k)\frac{\partial}{\partial u} f(\boldsymbol{r}_k, u_{k,b})\delta u_k = R_k \tag{22}$$

and $\delta u_k = b_k$ for all $k \in S_D$, then the following stability inequality holds for the discrete solution:

$$\max_{1\le k\le n}\left|\delta u_k\right| \le \max_{k\in S_D}\left|b_k\right| + \max_{k\in S_B}\frac{\left|R_k\right|}{\nu(B_k)\frac{\partial}{\partial u} f(\boldsymbol{r}_k, u_{k,b})} \tag{23}$$

This inequality is directly applicable for the estimation of the maximum Newton correction if the Jacobian on the left hand side of Equation (22) is used to solve the discretized nonlinear problem (19) by Newton's method. Moreover such stability inequalities are very useful for evaluating upper bounds for the discretization error even for the nonlinear problem (19). In Reference [25] an excellent example is given demonstrating clearly how bad the discretization error is controlled if the off diagonal elements of the Jacobian of the discretized electron continuity equation have both signs. Moreover it is shown as well in Reference [25] for the same problem that the discrete solution gets much more accurate and very well controlled by the applied voltages if all off diagonal elements of the Jacobian become always negative or zero after a modification of the grid. The underlying reason is not as falsely stated in Reference [21] that the original grid had one obtuse triangle but that the original grid was not the Delaunay grid constructed on the basis of the Voronoi diagram, whereas the modified grid is the Delaunay grid. As pointed out already earlier, a Delaunay grid may contain obtuse triangles!

Another advantage of the discretization scheme presented above is that it allows a straight forward incorporation of the Scharfetter-Gummel discretization formula for the balance equations, which was originally developed for rectangular grids [2, 4, 5]. The application of this discretization formula is mandatory for achieving accurate simulation results on coarse grids. For general finite element schemes it is typically very difficult to incorporate this formula but for the scheme described above it is easily done by choosing $\bar{a}(l_{k,j})$ as follows:

For holes:

$$\bar{a}(l_{k,j}) = n_{\mathrm{i}}\bar{\mu}_p(l_{k,j})V_{\mathrm{T}}\frac{\Psi_j - \Psi_k}{\exp(\frac{\Psi_j}{V_{\mathrm{T}}}) - \exp(\frac{\Psi_k}{V_{\mathrm{T}}})} \tag{24}$$

For electrons:

$$\bar{a}(l_{k,j}) = n_{\mathrm{i}}\bar{\mu}_n(l_{k,j})V_{\mathrm{T}}\frac{\Psi_k - \Psi_j}{\exp(\frac{-\Psi_j}{V_{\mathrm{T}}}) - \exp(\frac{-\Psi_k}{V_{\mathrm{T}}})} \tag{25}$$

$\overline{\mu_{n,p}}(l_{k,j})$ are appropriate mean values of the mobilities that can be regarded as a function of the edge $l_{k,j}$. Results concerning the consistency and convergence of the discretization scheme described above for the balance equations can be found in Reference [31].

Gummel's nonlinear relaxation method (4) performed in such a manner that the relevant discrete Jacobian matrices fulfill conditions **II**-**IV**, which typically means that derivatives with respect to the solution variables are neglected for the mobilities and impact ionization, converges nearly always even for very bad initial solutions. One of the rare counter examples is given in Reference [14]. The convergence is typically slow for high current applications but is very predictable so that the solution accuracy can be estimated very reliably during the iteration process [37]. This allows to switch to more coupled methods like a simultaneous Newton method not before the solution accuracy is so high that the simultaneous Newton method is in the range where it converges quadratically. Of course in this case all derivatives should be considered in the Jacobian matrix of the fully simultaneous Newton method. The availability of this combination of solution methods for the DD model featuring high robustness even for bad initial solution and high accuracy at the same time is possibly one reason why the DD model is still the numerical device model that is applied by far most even for nanoscale devices, where its physical accuracy is certainly questionable [38]. The above comments concerning the beneficial effect on robustness of neglecting the derivatives of the mobilities and impact ionization apply as well to other nonlinear relaxation methods [39, 14] and are even valid for the fully coupled Newton method outside the range of quadratic convergence. It is rather straight forward to extend the box integration method including a Scharfetter-Gummel type discretization for the energy flux densities to the HD model [6]. A nonlinear relaxation method with comparable convergence properties to Gummel's method has been published in Reference [40] and evaluated in Reference [41] for the HD model. For this method and the numerical solution algorithms of the HD model in general convergence robustness increases as well decisively if certain derivatives with respect to mobilities and impact ionization are turned off in order to enhance the "stability" of the discretized equations.

3. Conclusion

Based on the historic development of the space discretization and solution methods for the Drift Diffusion model it is shown how important it is for the error control on coarse grids to preserve especially the monotonicity properties of the underlying partial differential equations in the final discretized model. The author believes that this observation should serve as a guideline for the development of discretization methods for transport models in future.

References

[1] H. K. Gummel: A self-consistent iterative scheme for one-dimensional steady state transistor calculations, IEEE Transactions on Electron Devices, 455–465 (October 1964).

[2] D. L. Scharfetter, H. K. Gummel: Large-Signal Analysis of a Silicon Read Diode Oscillator, IEEE Transactions on Electron Devices **16**(1), 64–77 (1969)

[3] B. Meinerzhagen: *Effiziente Rechenstrategien zur Bauelementsimulation*, Doctoral Thesis (RWTH-Aachen, 1985)

[4] W. L. Engl, H. K. Dirks, B. Meinerzhagen: Device Modeling, Proc. of the IEEE **71**, 10–33 (1983)

[5] R. E. Bank, D. J. Rose, W. Fichtner: Numerical Methods for Semiconductor Device Simulation, IEEE Trans. Electron Devices **30**(9), 1031–1041 (1983)

[6] B. Meinerzhagen, W. L. Engl: The Influence of the Thermal Equilibrium Approximation on the Accuracy of Classical Two-Dimensional Numerical Modeling of Silicon Submicrometer MOS Transistors, IEEE Trans. Electron Devices **35**(5), 689–697 (1988)

[7] J. W. Slotboom: Computer-Aided Two-Dimensional Analysis of Bipolar Transistors, IEEE Trans. Electron Devices **20**, 669–679 (1973)

[8] C. Jungemann, B. Meinerzhagen: *Hierarchical Device Simulation: The Monte-Carlo Perspective*, Computational Microelectronics, ed. by S. Selberherr (Springer, Wien, New York 2003)

[9] S. Selberherr: *Analysis and Simulation of Semiconductor Devices* (Springer, Wien 1984)

[10] O. A. Ladyshenskaya, N. N. Uraltseva: *Linear and Quasi-linear Elliptic Equations* (Academic Press, New York 1968)

[11] M. S. Mock: *Analysis of Mathematical Models of Semiconductor Devices* (Boole Press, Dublin 1983)

[12] T. I. Seidmann: Steady State Solution of Diffusion Reaction Systems with Electrostatic Convection, Nonlinear Analysis Theory, Methods & Applications **4**, 623–637 (1979)

[13] J. W. Slotboom: Iterative scheme for 1 and 2-dimensional DC transistor simulations, Electron. Lett. **5**, 677–678 (1969)

[14] K. H. Bach, H. K. Dirks, B. Meinerzhagen, W. L. Engl: A New Nonlinear Relaxation Scheme for Solving Semiconductor Device Equations, IEEE Trans. Computer–Aided Des. **10**(9), 1175–1186 (1991)

[15] L. Collartz: *Funktionalanalysis und numerische Mathematik* (Springer-Verlag, Berlin Heidelberg New-York 1968)

[16] H. H. Heimeier: A Two-Dimensional Numerical Analysis of a Silicon N-P-N Transistor, IEEE Trans. Electron Devices **20**, 708–714 (1973)

[17] J. J. Barnes, R. J. Lomax: Finite-Element Methods in Semiconductor Device Simulation, IEEE Trans. Electron Devices **24**, 1082–1089 (1977)

[18] E. Buturla, P. E. Cottrell, B. Grossmann, K. Salsburg: Finite-Element Analysis of Semiconductor Devices: The Fielday Program, IBM J. Res. Develop. **25**, 218–231 (1981)

[19] G. D. Hachtel, M. H. Mack, R. O'Brian, B. Speelpenning: Semiconductor Analysis using Finite-Elements - Part I: Computational Aspects, IBM J. Res. Develop. **25**, 232–245 (1981)

[20] G. D. Hachtel, M. H. Mack, R. O'Brian: Semiconductor Analysis using Finite-Elements - Part I: IGFET and BJT Case Studies, IBM J. Res. Develop. **25**, 246–260 (1981)

[21] J. F. Burgler, R. E. Bank, W. Fichtner, R. K. Smith: A new Discretization Scheme for the Semiconductor Current Continuity Equation, IEEE Trans. on Computer-Aided Design **8**(5), 479–489 (1989)

[22] P. J. Mole: Discretization of the semiconductor current continuity equation for finite element solvers in 2 and 3 dimensions. In: *Proc. Fouth Conf. Numerical Analysis of Semiconductor Devices and Integrated Circuits-NASECODE IV*, ed. by J. J. H. Miller (Boole Press, Trinity College, Dublin, Irland 1985) pp. 429–435

[23] W. G. Strang, G. J. Fix: *An analysis of the finite element method* (Prentice-Hall, Englewood Cliffs 1973)

[24] G. Hachtel, M. Mack, R. O'Brien: Semiconductor Device Analysis Via Finite Elements. In: *Proc. of the Eighth Asilomar Conference on Circuits and Systems* (1974) pp. 332–338

[25] C. S. Rafferty, M. R. Pinto, R. W. Dutton: Iterative methods in semiconductor device simulation, IEEE Transactions on Electron Devices **32**(10), 2018–2027 (1985)

[26] R. S. Varga: *Matrix Iterative Analysis*, Series in Automatic Computation (Prentice-Hall, Englewood Cliffs, New Jersey 1962)

[27] R. H. MacNeal: An Asymmetrical Finite Difference Network, Quart. Apll. Math. **11**, 295–310 (1953)

[28] C. Großmann, H. G. Roos: *Numerische Behandlung partieller Differentialgleichungen* (B. G. Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden 2005)

[29] P. Cottrell, E. Buturla: Steady State Analysis of Field Effect Transistors via the Finite Element Method. In: *IEDM Tech. Digest* (IEEE, 1975) pp. 51–54

[30] E. B. D. C. Cole, S. Furkay, K. Varahramyan, J. Slinkman, J. Mandelman, D. Fotyi, O. Bula, A. Strong, J. Park, T. L. Jr, J. Johnson, M. Fischetti, S. Laux, P. Cottrell, H. Lustig, F. Pileggi, D. Katcoff: The Use of Simulation in Semiconductor Technology Development, Solid State Electronics **33**, 591–623 (1990)

[31] M. Sever: Delaunay Partitioning In Three Dimensions and Semiconductor Models, Compel **5**, 75–93 (1986)

[32] G. Voronoi: Nouvelles application des parametres continus a la theorie des forme quadratiques. Deuxieme memoire. Recherche sur les parallelloedres primitifs, Journal für reine und angewandte Mathematik **134**, 198–287 (1908)

[33] B. N. Delone: Sur la sphère vide, Bulletin of the Academy of Sciences of the U.S.S.R. **7**, 793–800 (1934)

[34] B. S. Baker, E. Grosse, C. S. Rafferty: Nonobtuse triangulation of polygons, Discrete & Computational Geometry **3**, 147–168 (1988)

[35] J. A. Meijerink, H. A. vanderVorst: An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, Math. Comp. **31**, 148–162 (1977)

[36] A. A. Samarskij: *Theorie der Differenzenverfahren* (Akademische Verlagsgesellschaft Geest & Portig K.-G., Leibzig 1984)

[37] K. Bach: Nichtlineare Relaxationsverfahren zur Lösung von Differentialgleichungssystemen am Beispiel der Halbleitergleichungen, Doctoral Thesis (RWTH-Aachen, 1991)

[38] C. Jungemann, T. Grasser, B. Neinhuis, B. Meinerzhagen: Failure of Moment-Based Transport Models in Nanoscale Devices near Equilibrium, IEEE Trans. Electron Devices **52**(11), 2404– 2408 (2005)

[39] B. Meinerzhagen, H. K. Dirks, W. L. Engl: Quasi-simultaneous solution method: A new highly efficient strategy for numerical MOST simulations, IEEE Trans. Computer–Aided Des. **4**, 575– 582 (1985)

[40] B. Meinerzhagen, K. Bach, I. Borg, W. L. Engl: A New Highly Efficient Nonlinear Relaxation Scheme for Hydrodynamic MOS Simulations. In: *NUPAD IV Technical Digest* (Seattle 1992) pp. 91–96

[41] Y. Apanovich, E. Lyumskis, B. Polsky, P. Blakey: An Investigation of Coupled and Decoupled Iterative Algorithms for Energy Balance Calculations. In: *Simulation of Semiconductors and Processes*, ed. by S. Selberherr, H. Stippel, E. Strasser (Springer Verlag, Wien 1993) pp. 233–236

Photography & Biography

Bernd Meinerzhagen is Full Professor and Head of the Institute for Electron Devices and Circuits at the Technical University Braunschweig, Braunschweig, Germany. He received the Dipl.-Ing. degree in electrical engineering, the Dipl.-Math. degree in mathematics, the Dr.-Ing. degree in electrical engineering, and the “venia legendi,” all from the RWTH Aachen (Aachen University of Technology), Aachen, Germany, in 1977, 1981, 1985, and 1995, respectively. From 1978 to 1986, he worked mainly on the development of numerical device modeling codes as a Research and Teaching Assistant at the RWTH Aachen. In 1986, he joint AT&T Bell Laboratories, Allentown, PA, as a Member of Technical Staff, where he developed advanced numerical models for MOS substrate and gate currents. From 1988 to 1995 he was the Head of the Research and Development Group for Silicon technology modeling and simulation (TCAD) at the RWTH Aachen and he continued this research as Professor at the University of Bremen, Bremen, Germany between 1995 and 2003. His research interests are focused on the physics, characterization, modeling and design of Si/SiGe integrated devices and circuits and on the mathematical foundations of electrical engineering. Prof. Meinerzhagen has coauthored a book and more than 200 papers published in international journals and conference proceedings and has been a Technical Program Committee member of IEDM, ESSDERC, SISPAD, IWCE and other conferences.

J. Microelectron. Manuf. **3**, 20030404 (2020)
doi: 10.33079/jomm.20030404

Nano-Electronic Simulation Software (NESS): A Novel Open-Source TCAD Simulation Environment

Cristina Medina-Bailon [*], Tapas Dutta, Fikru Adamu-Lema, Ali Rezaei, Daniel Nagy, Vihar P. Georgiev[**], and Asen Asenov

Device Modelling Group, James Watt School of Engineering, University of Glasgow, G12 8LT Glasgow, UK.

Abstract: This paper presents the latest status of the open source advanced TCAD simulator called Nano-Electronic Simulation Software (NESS) which is currently under development at the Device Modeling Group of the University of Glasgow. NESS is designed with the main aim to provide an open, flexible, and easy to use simulation environment where users are able not only to perform numerical simulations but also to develop and implement new simulation methods and models. Currently, NESS is organized into two main components: the structure generator and a collection of different numerical solvers; which are linked to supporting components such as an effective mass extractor and materials database. This paper gives a brief overview of each of the components by describing their main capabilities, structure, and theory behind each one of them. Moreover, to illustrate the capabilities of each component, here we have given examples considering various device structures, architectures, materials, etc. at multiple simulation conditions. We expect that NESS will prove to be a great tool for both conventional as well as exploratory device research programs and projects.

Keywords: Integrated Simulation Environment, Variability, Drift-Diffusion, Quantum Correction, Kubo-Greenwood, Non-Equilibrium Green's Function.

1. Introduction

Two of the major issues with experimental research and design are cost and time. Technology computer-aided design (TCAD) plays a crucial role in reducing the development costs and time-to-market for the semiconductor industry by performing physical analysis of already existing devices or novel technologies and transistor architectures [1]. Therefore, in the development of the TCAD tools, there are two key objectives: accurate physical models and reduced simulation time.

A great amount of commercially available TCAD software [2,3] as well as academic simulation tools with different levels of complexity, including drift-diffusion (DD) with quantum corrections [4,5], 3D ensemble Monte Carlo (MC) [6-8], multi-subband (MS) 2D [9] and 1D MC [10], direct Boltzmann Transport Equation (BTE) solvers [11], Non-Equilibrium Green's Function (NEGF) simulators in ballistic regime [12] and with scattering [13] already exist. However, the commercial TCAD tools so far are not open source software, which limits collaboration. Meanwhile, the academic software tends to work in isolation, and it is difficult to investigate a particular transistor structure with different complexity of simulation techniques [14].

In this paper, we introduce the concepts and the inner workings of a user-friendly and open-source TCAD semiconductor device simulator called Nano-Electronic Simulation Software (NESS), developed by the Device Modelling Group at the University of Glasgow. NESS enables simulations, with increasing complexity and physical content within a unified environment. Open source also means that it allows collaboration and co-development by industry and academia all over the world. NESS is designed to be flexible, easy to use, and extendable thanks to its modular structure [14].

This paper is organized as follows. In Section 2, we provide a brief overview of the NESS structure. In Section 3, we discuss the structure generator (SG) module used for the generation of the mesh and device structure which are used as an input file for rest components of NESS. Section 4 provides a detailed overview of the numerical modules already implemented: DD, Kubo-Greenwood (KG) and, NEGF. Finally, in Section 5, we finish with the concluding remarks.

[*] Address all correspondence to Cristina Medina-Bailon, E-mail: cristina.medinabailon@glasgow.ac.uk

[**] Address all correspondence to Vihar P. Georgiev, E-mail: vihar.georgiev@glasgow.ac.uk

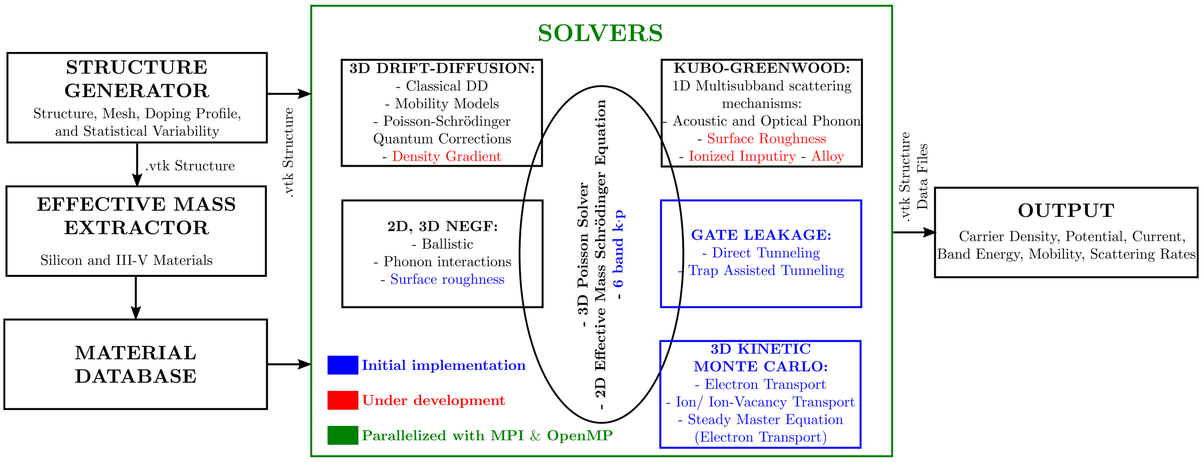

STRUCTURE GENERATOR
Structure, Mesh, Doping Profile, and Statistical Variability
.vtk Structure
EFFECTIVE MASS EXTRACTOR
Silicon and III-V Materials
MATERIAL DATABASE
.vtk Structure
SOLVERS
3D DRIFT-DIFFUSION:
- Classical DD
- Mobility Models
- Poisson-Schrödinger Quantum Corrections
- Density Gradient
2D, 3D NEGF:
- Ballistic
- Phonon interactions
- Surface roughness
- 3D Poisson Solver
- 2D Effective Mass Schrödinger Equation
- 6 band k·p
KUBO-GREENWOOD:
1D Multisubband scattering mechanisms:
- Acoustic and Optical Phonon
- Surface Roughness
- Ionized Imputiry - Alloy
GATE LEAKAGE:
- Direct Tunneling
- Trap Assisted Tunneling
3D KINETIC MONTE CARLO:
- Electron Transport
- Ion/ Ion-Vacancy Transport
- Steady Master Equation (Electron Transport)
Initial implementation
Under development
Parallelized with MPI & OpenMP
.vtk Structure
Data Files
OUTPUT
Carrier Density, Potential, Current, Band Energy, Mobility, Scattering Rates

Figure 1. Flowchart of NESS detailing its modular structure.

2. Overview of NESS

In this section, we provide an overview of our simulation environment NESS and its modular structure. Currently, there are five main components of NESS which are summarized in Figure 1: SG, effective mass extractor, material database, solvers, and outputs. First, the SG [15,16] (more details in Section 3) is used to generate and configure the 3D device structures such as nanowires (NWs), multi-gate 3D device architectures, or bulk complementary metal-oxide-semiconductor (CMOS) transistors. It allows users to consider different semiconductor materials (such as Si, Ge, or III-Vs materials), doping configurations (such as uniform or Gaussian profiles), mesh designs, and the main variability sources (random discrete dopants (RDD), line edge roughness (LER), and metal gate granularity (MGG)).

Second, as the effective masses strongly depend on the confinement orientation of the nanostructures, an automated routine to extract the effective mass from first principle simulations has been implemented in NESS [1]. It can calculate the correct electron confinement and transport effective masses from atomistic simulations (such as Density Functional Theory (DFT)) or semi-empirical models (such as Tight-Binding (TB)) of the electronic band structure of NW with the technologically relevant cross-sectional area, shape, and transport orientations.

Third, the material database provides the relevant parameters for each material considered in the generated structure, such as the work-function, affinity, dielectric constants, mobility model parameters, or scattering parameters. Furthermore, the effective masses can be provided for each material from DFT and TB methods, or directly from our effective mass extractor. As illustrated in Figure 1, those parameters serve as input for the solvers.

Fourth, different transport simulation solvers [14] have been already implemented in NESS to simulate the mobility, the charge density, and the current in nano-CMOS devices. They have been implemented with a high degree of parallelism making use of MPI and OpenMP libraries. In general, each of them is self-consistently solved with the 3D Poisson and/or the 2D Schrödinger equations. Section 4 describes in details the three current main numerical solvers: *(i)* DD module, which contains different mobility models and Poisson-Schrödinger quantum corrections [17]; *(ii)* KG module, which calculates the low-field electron mobility; and *(iii)* the coupled mode-space NEGF solver, which captures quantum mechanical effects, coherent transport, and impact of scattering. Moreover, different enhanced modules and solvers [18] are currently under development in NESS including: density gradient; extension of the KG module [19] to consider surface roughness (SR), ionized impurity, and alloy scattering mechanisms; implementation of SR scattering mechanism in the existing NEGF module [20]; Kinetic MC solver [21] for the simulation of memory devices; module to compute the gate leakage current; and a full-band quantum transport solver in presence of hole-phonon interactions using a mode-space k·p approach in combination with the existing NEGF module [22].

Finally, the simulation results (i.e. current, electrostatic potential, charge concentration) are stored in text files and in vtk format for easy visualization with freeware software, such as ParaView.

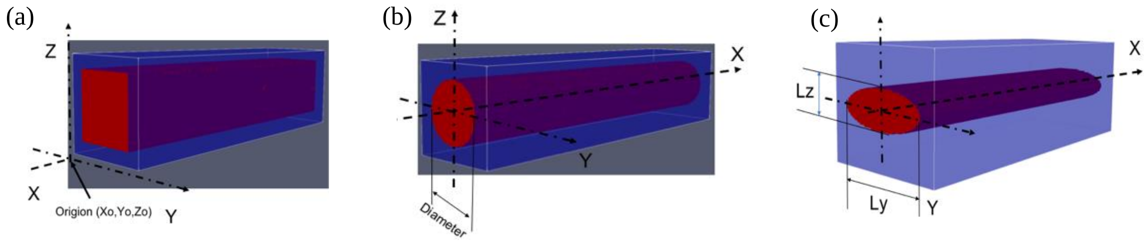


Figure 2. Some of the main primitive objects that can be used to create complex device structures.

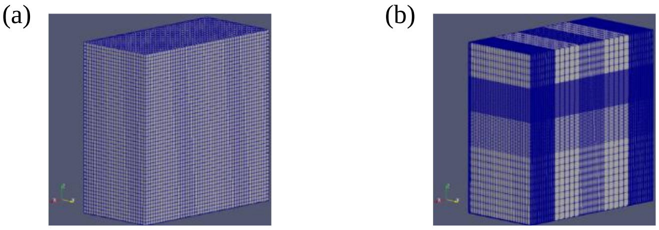


Figure 3. Uniform and non-uniform mesh generation examples.

3. Structure Generator

In this section, we introduce the device SG and provide some examples. The SG is a flexible module capable of generating various types of devices and the corresponding structures (simulation domains). The generated device structure data file can be stored as a binary or ASCII format where the datasets are defined by the rectilinear grid with a regular topology along the coordinates.

Creation of geometric objects: Users can create any type of polygon shape and three main types of geometric objects, which are (a) cuboid, (b) cylinder with circular cross-section and (c) cylinder with elliptical cross-section as shown in Figure 2. The simple elliptical shape ($z^2/l_z + y^2/l_y$) assumes that the origin is located at (0,0,0), and implemented in NESS to create both cylinder types. When assigning material and doping properties to the mesh, NESS makes two important assumptions. Materials are considered as a property of an element defined by a volume of ($\Delta V = \Delta x \cdot \Delta y \cdot \Delta z$). On the other hand, doping is assigned to a discretization node. Users can generate uniform (Figure 3(a)) and non-uniform (Figure 3(b)) meshes for their device structure.

Bulk MOSFET and SOI example: Figure 4 shows examples of conventional bulk MOSFET and fully depleted Silicon on Insulator (FDSOI) structures, generated using the NESS SG.

Statistical variability: The contemporary CMOS transistors are highly susceptible to statistical variability and their performance and electrical characteristics could be significantly affected by it. The SG can introduce the main sources of statistical variability in the device structure prior to running the simulations. In NESS, users can choose from three sources of variability: RDD [23], LER [24], and MGG [25]; or they can run simulations considering all sources of variability or different combinations of them. Figure 5 shows a randomly generated atomistic device considering RDD and MGG in the simulation domain.

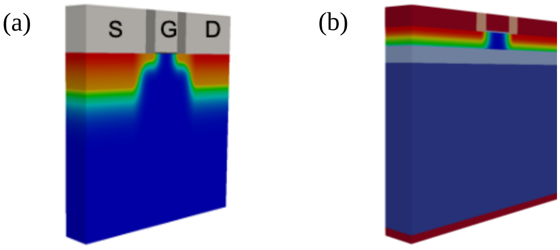


Figure 4. (a) Conventional bulk MOSFET, and (b) FDSOI.

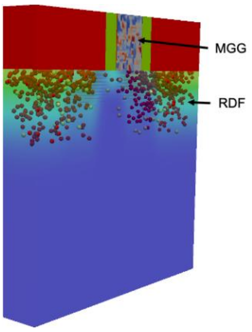


Figure 5. Atomistic device considering RDD and MGG.

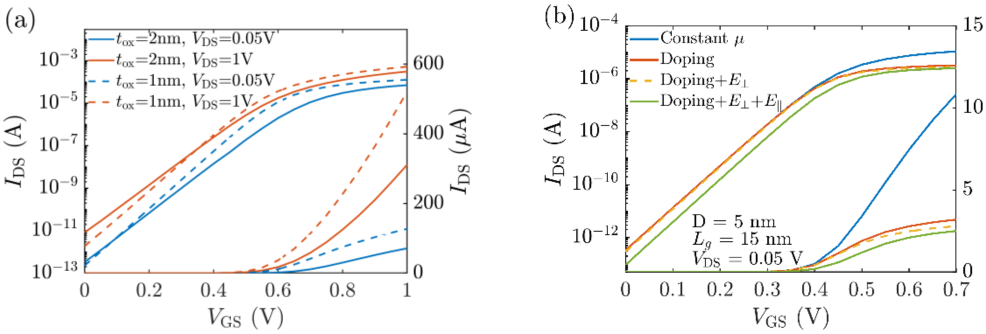


Figure 6. (a) Transfer characteristics of the bulk MOSFET shown in Figure 4(a) at low and high drain bias for t_{ox}=1nm, 2nm using DD. Constant bulk mobility of 1400 $cm^2V^{-1}s^{-1}$ was used. (b) Impact of mobility models on the transfer characteristics of a nanowire FET with circular cross section having a diameter of 5nm and channel length of 15nm, $N_{Channel}$=10^{15} cm^{-3} and N_{SD}=10^{20} cm^{-3}. The low field mobility used was 481 $cm^2V^{-1}s^{-1}$ calculated using the KG module including the impact of acoustic and optical phonon scattering mechanisms (Section 4.2).

4. Numerical Solvers

4.1 Drift-Diffusion

The DD formalism for carrier transport has been the main workhorse in the TCAD industry for many decades. It is indispensable for simulating bulk CMOS transistors and relatively larger devices where a more sophisticated approach is neither desired nor practical.

In NESS, we have implemented the DD module using a finite volume discretization scheme for the current continuity equation following the Scharfetter-Gummel approach [26] using Bernoulli functions. The 3D current continuity equation is self-consistently solved with the 3D Poisson equation until convergence. Different mobility models are included in the current continuity equation. Convergence for potential and charge is reached when the max norm of the difference between two successive Gummel iterations reaches the preset criteria. At present, we have included doping dependence of the mobility using the Masetti model [27]. The transverse and longitudinal electric field ($E_\perp$, $E_\parallel$, respectively) dependence of the mobility has been included by means of the Yamaguchi model [28] and the Caughey-Thomas [29] model, respectively. As examples, simulation results for a conventional bulk MOSFET with channel length of 25nm for two oxide thicknesses are shown in Figure 6(a), for low and high drain bias conditions considering constant bulk mobility. In Figure 6(b), we have shown the cumulative impact of the mobility models on the transfer characteristics for a nanowire transistor with a circular cross-section of 5nm diameter and 15nm channel length.

A key issue with classical DD simulations is that they cannot capture the quantum confinement effects. A quantum-corrected DD simulator can ensure a correct charge profile in the device at a fraction of the computational cost of a full quantum simulator. We have developed and implemented Schrödinger equation-based quantum-corrected DD approach in NESS [17]. For this, we first self-consistently solve the 2D Schrödinger equation in planes perpendicular to transport and 3D Poisson equation in the whole device. The 3D quantum charge is calculated using a top of the barrier approach [30], summing over all subbands and valleys. At convergence, the quantum charge density (n_Q) is used to calculate a quantum correction term $k_BT/q\log(n_Q/N_C)$ where N_C is the conduction band density of states, T is temperature, k_B is Boltzmann constant, and q is the electronic charge [31,32]. This term is then used to generate a corrected potential which (instead of the classical potential obtained from the Poisson equation) is used as a driving force in the continuity equation. This is repeated until the charge and the potential converge. The quantum correction can either be fixed for a bias point (for low drain voltage) or can be updated in each Gummel iteration. It is worthwhile to note that this approach does not use any fitting parameters in the quantum correction procedure unlike the density gradient or the effective potential method.

The quantum-corrected DD remedies the deficiency of the classical DD charge profile as can be seen in Figure 7(a) for a nanowire FET with an 5nm × 5nm square cross-section. Further, in contrast to the classical DD, current-voltage characteristics obtained using quantum-corrected DD display the shift in threshold voltage due to quantum confinement with is in an excellent match with the result obtained from ballistic NEGF as shown in Figure 7(b).

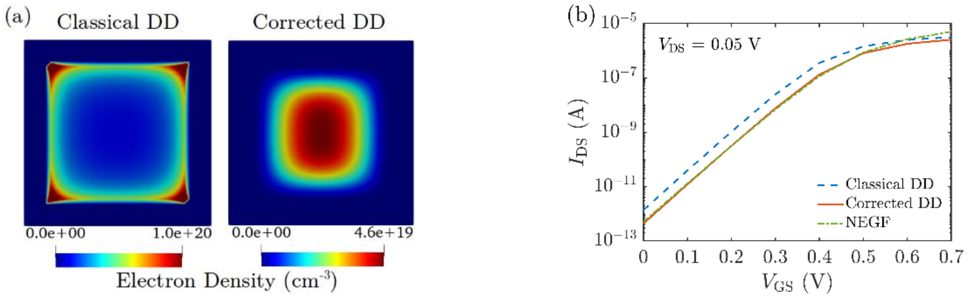


Figure 7. (a) 2D profile of electron density in a [110] oriented NW with 5nm × 5nm square cross section and L_G=10nm for classical (left) and quantum-corrected (right) DD in the plane normal to transport direction at the middle of the channel at V_{GS}=0.7V, V_{DS}=0.05V. (b) I_{DS}−V_{GS} characteristics calculated using classical DD, quantum-corrected DD, and ballistic NEGF. Note that in these simulations, in case of NEGF and for charge calculation after solving Schrödinger equation in corrected DD, the Fermi level at the source, E_{FS} is set to the quasi-Fermi level at the source contact as obtained in DD. The low field mobility used was 477 cm^{2}V^{-1}s^{-1} calculated using the KG module.

4.2. Kubo-Greenwood Module

The KG solver implemented in NESS provides accurate electron mobility at low-field near-equilibrium conditions [33,34]. It combines the quantum effects based on the 1D multi-subband scattering rates of the most relevant scattering mechanisms in confined channels [19] and the semi-classical BTE by applying the KG formula within the relaxation time approximation [11]. In the first step, the NEGF module of NESS is used to extract the electron densities, subband levels (E_l), and the corresponding wavefunctions (ξ_l) at the cross-section area of a gated NW in the presence of a low electric field in the transport direction (the long-channel device approximation).

In the second step, the 1D rates for the dominant scattering mechanisms in silicon are calculated using the parameters from the first step. The scattering rates are directly derived from the Fermi Golden Rule, using the time-dependent perturbation theory and assuming that the transitions between two states occur instantaneously. In this paper, we present two of the implemented scattering mechanisms:

Acoustic (Ac) phonon scattering is considered to be elastic and within the short-wave vector limit. Its equivalent equation from an initial subband l and a final subband l' is:

$$\Gamma_{Ac}(l,k)=\frac{|D_{Ac}|^2 k_B T}{\rho\hbar u_s^2}\frac{m_v}{\hbar^2}\sum\nolimits_{l'}\left[\int d\vec{s}\,|\xi_l(\vec{s})|^2|\xi_{l'}(\vec{s})|^2\right]\times\theta\left(\epsilon(k)+\Delta E_{l'}\right)\left(\frac{1}{|q_1+k|}+\frac{1}{|q_2+k|}\right), \quad (1)$$

where D_{Ac} is the acoustic deformation potential, ρ is the material density, $\hbar$ is the reduced Planck's constant, u_s is the speed of sound, m_l is the electron effective mass in the transport direction, $\vec{s}$ are vectors normal to the transport direction, θ represents the step function, $\epsilon(k)$ is the kinetic energy for a wavevector with magnitude k, $\Delta E_{l'}=E_{l'}-E_l$ is the energy separation between subbands l and l', and $q_{1/2}=-k\pm\sqrt{k^2+\frac{\Delta E_{l'}2m}{\hbar^2}}$.

Optical (Op) phonon scattering takes into account g-type and f-type transitions (intra- and inter-valley transitions, respectively) and the energies of the different branches of the optical deformation potential are considered constant (as used in most of the standard approaches). Accordingly, the optical phonon scattering rate for the phonon mode j can be written as:

$$\Gamma_{Op}(j,l,k)=\frac{|D_{Op,j}|^2}{2\rho\omega_j}\sum\nolimits_{l'}\left[\int d\vec{s}\,|\xi_l(\vec{s})|^2|\xi_{l'}(\vec{s})|^2\right]\times\int dqG(q), \quad (2)$$

where

$$\int dqG(q)=\frac{n_j\theta\left(\epsilon(k)+\Delta E_{l'j}^{+}\right)m_{v'}}{\hbar^2}\left(\frac{1}{|q_1+k|}+\frac{1}{|q_2+k|}\right)+\frac{\left(n_j+1\right)\theta\left(\epsilon(k)+\Delta E_{l'j}^{-}\right)m_{v'}}{\hbar^2}\left(\frac{1}{|q_3+k|}+\frac{1}{|q_4+k|}\right), \quad (3)$$

with

Table 1. Main dimensions, doping values, and scattering parameters for the cylindrical Si NWs.

Device Parameters	Si width	From 3nm to 6nm	Scattering parameters	D_{Ac}	14eV
	SiO2 width	0.8nm		$D_{Op,j}$ (g-type)	[5,8,30] ·10^9eV/m
	Doping	10^{15}cm^{-2}		$D_{Op,j}$ (f-type)	[1.5,34,40] ·10^9eV/m
	Temperature	300		ω_j (g-type)	[0.01206,0.01853,0.063] eV
	Effect. Mass	Ref. [1]		ω_j (f-type)	[0.01896,0.0474,0.05903] eV

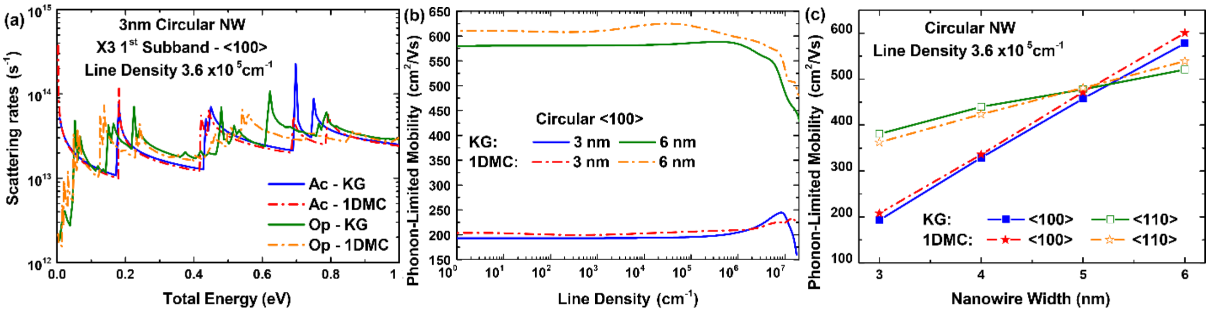


Figure 8. Scattering rates and mobility comparison between the KG module from NESS and the 1DMC code from [10]: (a) Acoustic and optical phonon scattering rates of the first subband of valley X3 as a function of the total energy for a 3nm circular NW with a line density of 3.6×10^{5}cm^{-1} and ⟨100⟩ orientation. (b) Phonon-limited electron mobility as a function of the line density for 3nm and 6nm circular NWs with ⟨100⟩ orientation. (c) Phonon-limited electron mobility as a function of the width for circular NWs with a line density of 3.6×10^{5}cm^{-1}, ⟨100⟩ and ⟨110⟩ orientations.

$$q_{1/2} = -k \pm \sqrt{\frac{m_{v'}}{m_v}k^2 + \frac{\Delta E^{+}_{l'j}2m_{v'}}{\hbar^2}};$$
$$q_{3/4} = -k \pm \sqrt{\frac{m_{v'}}{m_v}k^2 + \frac{\Delta E^{-}_{l'j}2m_{v'}}{\hbar^2}} \text{ and} \qquad (4)$$
$$\Delta E^{+/-}_{l'j} = E_{l'} - E_l \pm \hbar\omega_j,$$

Here, n_j is the phonon number, ω_j is the phonon energy, $D_{Op,j}$ is the optical deformation potential, and $m_v(m_{v'})$ is the transport effective mass of the initial(final) valleys, respectively.

In the third step, the mobility (μ_i^l) for the scattering mechanism *i* and subband *l* is calculated considering the semi-classical simulation of the transport properties of a 1D electron gas using the BTE within the relaxation time approximation [11] as a function of the relaxation time ($\tau_i^l(E) = 1/\Gamma_i^l(E)$), the 1D density of states (g_l), the Fermi-Dirac function (f_0), and the 1D electron concentration (N_l):

$$\mu_i^l = \frac{2q}{k_B T N_l m_l}\int dE g_l(E)(E - E_l)\tau_i^l(E) f_0(E)(1 - f_0(E)). \qquad (5)$$

In the fourth step, we calculate in two strategies the total mobility for the *l* subband (μ^l): *(1)* it is calculated as a function of the individual mobilities associated with each scattering mechanism (μ_i^l) using the Matthiessen rule ($1/\mu^l = \sum_i 1/\mu_i^l$); and *(2)* the scattering rates of all mechanisms are directly added to avoid the Matthiessen rule and thereby μ^l is computed using Equation (5). The former strategy is of special interest for devices with large cross sections because the error induced by the Matthiessen rule in narrower devices is comparable to MS-MC and NEGF approaches. Finally, the average mobility of a NW structure is calculated accounting for all the subbands: $\mu_{NW} = \sum_l N_l \mu^l / \sum_l N_l$. The advantage of both semi-classical alternatives in comparison to purely quantum transport simulations is that the rates are individually computed and then combined, reducing dramatically the computational cost.

Figure 8 shows the scattering rates and mobility for cylindrical Si NWs, which main parameters are summarized in Table 1. The results from the KG module have been compared to the results of an external to NESS 1DMC simulator [10], where the mobility is extracted after applying a small constant electric field by fitting the average velocity versus field dependence. In general, the 1DMC and KG scattering rates for the lowest subband of the 3nm nanowire (Figure 8(a)) are in very good agreement especially at low energy levels, the most relevant region which determines the accuracy of the low-field mobility calculations. Moreover, the phonon-limited mobility computed with both approaches shows a very good agreement as a function of the line density (Figure 8(b)) for a 3nm and 6nm circular NW with ⟨100⟩, and as a function of the NW widths (Figure 8(c)) at a fixed line density for ⟨100⟩ and ⟨110⟩ orientations

4.3. NEGF

The so-called NEGF formalism, which is derived based on the Keldysh technique [35], is a widely applied framework for analyzing the electronic transport in non-equilibrium many-body systems. This method allows a quantum treatment of charge transport in order to capture quantum phenomena such as tunneling, coherence, and particle-particle interactions in mesoscopic and nanoscale devices. We obtain the charge density, potential profile, and the current flow in the system by performing a self-consistent solution of the Poisson equation and the NEGF transport equations in coupled-mode space (CMS). We can either consider diffusive transport by switching on the acoustic- and/or optical-phonon scattering [36,37] to enable electron-phonon (e–ph) interactions within the self-consistent Born approximation (SCBA) or neglect them to investigate merely the ballistic transport [13]. Moreover, we can simulate 2D planar structures such as DGSOI [38], and the NEGF solver implemented in NESS also allows calculation of the band-to-band tunneling by using the Flietner model to compute the current in heterostructures with a direct bandgap [39].

Adopting the notation of Reference [14], we will summarize the main concepts required to understand the NEGF formalism. Having the system in a steady state, the retarded, advanced, and lesser/greater Green's functions in real space representation read:

$$G^R(E) = \frac{1}{(E+i\eta)\cdot I - h - \Sigma^R(E)}, G^A(E) = \left[G^R(E)\right]^\dagger, \tag{6}$$

$$G^{\lessgtr} = G^R(E)\cdot\Sigma^{\lessgtr}(E)\cdot G^A(E), \tag{7}$$

where, h, and $\Sigma^{R(\lessgtr)}$ represent the one-particle Hamiltonian, and the retarded (lesser/greater) self-energies accounting for electrons interactions with their surroundings, respectively. The charge at position r and the current take the forms:

$$n(r) = -\frac{i}{2\pi}\int dE\, G^<(r, r'; E) \tag{8}$$

$$j(l) = \frac{2\cdot|q|}{\hbar}\int\frac{dE}{2\pi}\mathrm{Tr}\left[2\,\mathrm{Re}\left\{h_{l+1,l}\cdot G^<_{l,l+1}\right\}\right] \tag{9}$$

Here $h_{l+1,l}(G^<_{l,l+1})$ indicates the matrix elements of the Hamiltonian (lesser Green's function) between the basis states in layer $l+1$ (l) and l ($l+1$) [12,40].

Before considering the e–ph interactions, let us briefly discuss the CMS approximation. The single-particle Hamiltonian in the EM approximation can be expressed as:

$$\begin{aligned} h(r) &= h_T + h_L \\ &= \left[-\frac{\hbar^2}{2m^*_{y,z}}\Delta_{y,z} + V(r)\right] - \frac{\hbar^2}{2m^*_{y,z}}\frac{\partial^2}{\partial x^2} \end{aligned} \tag{10}$$

We can obtain the CMS representation by projecting each diagonal block $h_{n,n}$ of h_T on a subspace spanned by some chosen eigenmodes $\phi_i(y,z;n)$ of $h_{n,n}$ [41]. The transformation matrix is unitary in the limit where all the transverse modes are selected and, consequently, the CMS Hamiltonian is exactly equivalent to the real space Hamiltonian. On the other hand, the CMS Hamiltonian with few chosen modes is equal to the full rank EM Hamiltonian on the chosen subspace, as it reproduces by construction the exact selected EM sub-bands and their wavefunctions. Therefore, CMS offers the possibility to reproduce the effect of roughness or ionized impurities if enough modes are chosen. In this approximation, the matrix elements between the modes i and j read

$$\begin{aligned} &\tilde{G}^{R,\lessgtr}(l,i;l',j;E) = \\ &\quad \Sigma_{y,z}\Sigma_{y',z'}\phi_i^*(y,z;l)\cdot G^{R,\lessgtr}(l,y,z;l',y',z';E)\cdot\phi_j(y',z';l') \end{aligned} \tag{11}$$

To study the diffusive transport, the interactions of the electrons with phonons is implemented within the NESS via introducing the corresponding self-energies in real space [42,43]:

$$\Sigma^<_{\mathrm{ac},v}(r;E) = M_{\mathrm{ac}}G^<_v(r;E), \tag{12}$$

$$\Sigma^{<(>)}_{\mathrm{op},v}(r;E) = \Sigma_{q,v'}\left|M_q^{v,v'}\right|^2\left[n_{B,q}\cdot G^{<(>)}_{v'}\left(r;E-(+)\hbar\omega_q\right) + \left(n_{B,q}+1\right)\cdot G^{<(>)}_{v'}\left(r;E+(-)\hbar\omega_q\right)\right], \tag{13}$$

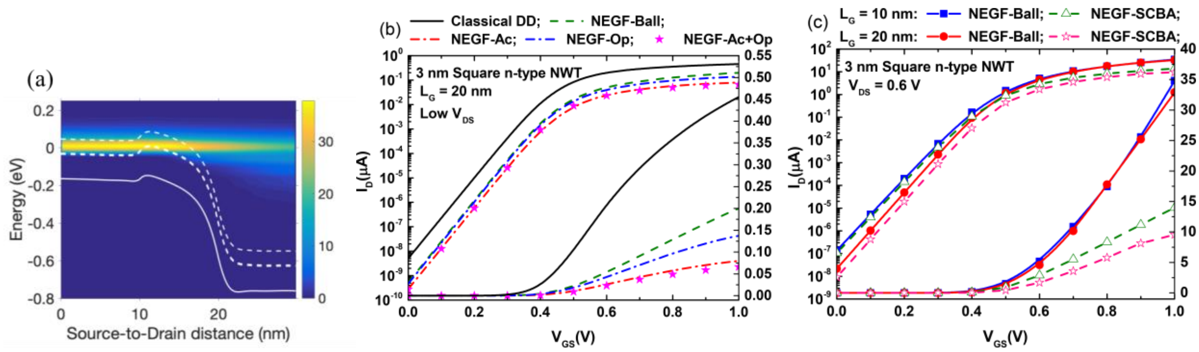


Figure 9. (a) The current spectrum in µA/eV for a NW with a square cross-section of 3nm×3nm and L_G=10nm calculated in the diffusive limit including e-ph scattering for ON-state (V_G=0.6V). The Fermi level at the source is the energy reference and V_{DS} = 0.6V. Moreover, the first subband of each valley is indicated with a white dashed line. The solid line corresponds to the potential along the transport direction which is the same as the first subband. I_{DS}−V_{GS} characteristics for n-type square 3nm×3nm Si NW assuming ballistic and dissipative NEGF transport simulations with: (b) L_G=20nm and low V_{DS} using the classical DD and the NEGF modules (including a combination of acoustic (Ac) and g-type optical (Op) phonon scattering mechanisms); and (c) L_G=10nm and L_G=20nm at V_{DS} = 0.6V.

where v, q, and $n_{B,q}$ refer to the electronic valley index, optical phonon with energy $\hbar\omega_q$, and the Bose-Einstein occupation number. The coupling constant of acoustic phonons M_{ac} , and the coupling strength of e–ph interaction $M_q^{v,v'}$ are obtained from the deformation potential theory [44]. The retarded component of the self-energy stemming from e-ph interactions may be expressed as:

$$\Sigma^{\mathrm{R}}(r;E)=\frac{1}{2}\left[\Sigma^{>}(r;E)-\Sigma^{<}(r;E)\right]. \tag{14}$$

Its CMS counterpart has the same form, whereas the real space self-energies are replaced by the CMS ones. Following the same notation as in Equation (11), and assuming that the self-energies are local in both space and time, the self-energies due to e-ph interactions in CMS representation read [45]:

$$\tilde{\Sigma}_{ac}^{<}(x,i;x,j;E)=M_{\mathrm{ac}}\sum\nolimits_{k,l}F_{k,l}^{i,j}(x)\tilde{G}^{<}(x,k;x,l;E) \tag{15}$$

$$\tilde{\Sigma}_{op,v}^{<(>)}(x,i;x,j;E)=\sum\nolimits_{k,l}F_{k,l}^{i,j}(x)\sum\nolimits_{q,v'}\left|M_q^{v,v'}\right|^2\left[\left(n_{B,q}+\frac{1}{2}\pm\frac{1}{2}\right)\tilde{G}_{v'}^{<(>)}\left(x,k;x,l;E\pm(\mp)\hbar\omega_q\right)\right]. \tag{16}$$

We can define the total retarded (lesser) self-energy as $\Sigma^{R(<)}=\Sigma_C^{R(<)}+\Sigma_{\mathrm{Scat}}^{R(<)}$, where $\Sigma^{R(<)}$ refers to the impact of electron exchange with the contacts [14].

In Figure 9(a), we show the ON-state-current spectrum resulting from the simulations for a 3nm × 3nm square NW transistor including scattering processes at V_{DS} = 0.6V. The tunnelling current reaches high values up to 30 μA/eV. Overall current damping is observed due to acoustic phonons and energy relaxation of carriers as they approach the drain due to optical phonons emission. Figures 9(b) and (c) show the I_{DS}−V_{GS} characteristics for a n-type 3nm×3nm square Si NW assuming ballistic and dissipative NEGF transport simulations. Figure 9(b) shows the results with L_G=20nm using the classical DD and the NEGF modules (Ac, g-type optical Op, and a combination of both phonon scattering mechanisms) at low drain voltage, whereas Figure 9(c) compares the results with L_G=10nm and L_G=20nm at V_{DS}=0.6V. More results from the NEGF module of NESS are presented in [1,14,15,16,18,20,22,39].

5. Conclusion

In this paper, we have described the organization and features of NESS - the new state-of-the-art device simulator from the Device Modeling Group at the University of Glasgow. We have highlighted the philosophy behind the project and demonstrated the capabilities of the various modules that are ready for release as open-source software. NESS encompasses everything that is required for modern nanodevice simulation – a tool for structure generation, effective mass extractor, low-field mobility simulator, and a large array of carrier transport solvers – ranging from classical to semi-classical and quantum formalisms. There are several new modules under active development. We hope the device community will find NESS useful for advanced device research.

Acknowledgments

This project was initiated by the European Union Horizon 2020 research and innovation programme under grant agreement No. 688101 SUPERAID7 and has received further funding from EPSRC UKRI Innovation Fellowship scheme under grant agreement No. EP/S001131/1 (QSEE), No. EP/P009972/1 (QUANTDEVMOD) and No. EP/S000259/1(Variability PDK for design based research on FPGA/neuro computing); and from H2020-FETOPEN-2019 scheme under grant agreement No.862539-Electromed-FET OPEN. The coauthors would like to thank Dr. Carrillo-Nuñez, Dr. Lee, Dr. Berrada, Dr. Badami, and Dr. Duan for their former contribution to NESS; as well as Dr. Donetti for the possibility of using the 1DMC tool.

References

[1] O. Badami, C. Medina-Bailon, S. Berrada, *et al.*, "Comprehensive Study of Cross-Section Dependent Effective Masses for Silicon Based Gate-All-Around Transistors," *Applied Sciences* **9**(9), 1895, 2019.

[2] Synopsys inc, 2017, [online] Available: https://solvnet.synopsys.com/

[3] Silvaco, [online] Available: https://silvaco.com/

[4] S. Selberherr, "Simulation of Semiconductor Devices and Processes," *Springer-Verlag*, 1993.

[5] A. J. Garcia-Loureiro, N. Seoane, M. Aldegunde, *et al.*, "Implementation of the Density Gradient Quantum Corrections for 3-D Simulations of Multigate Nanoscaled Transistors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **30**(6), 841-851, 2011.

[6] B. Winstead and U. Ravaioli, "A quantum correction based on Schrodinger equation applied to Monte Carlo device simulation," *IEEE Transactions on Electron Devices* **50**(2), 440-446, 2003.

[7] C. Riddet, A.R. Brown, C.L. Alexander, *et al.*, "3-D Monte Carlo Simulation of the Impact of Quantum Confinement Scattering on the Magnitude of Current Fluctuations in Double Gate MOSFETs," *IEEE Transactions on Nanotechnology* **6**, 48, 2007.

[8] N. Seoane, D. Nagy, G. Indalecio, *et.al.*, "A Multi-Method Simulation Toolbox to Study Performance and Variability of Nanowire FETs," *Materials* **12**(15), 2391, 2019.

[9] C. Medina-Bailon, J.L. Padilla, T. Sadi, *et al.*, " Multisubband ensemble Monte Carlo analysis of tunneling leakage mechanisms in ultrascaled FDSOI, DGSOI, and FinFET devices," *IEEE Trans. Electron Devices* **66**(3), 1145-1152, 2019.

[10] L. Donetti, C. Sampedro, F.G. Ruiz, *et al.*, "Multi-Subband Ensemble Monte Carlo simulations of scaled GAA MOSFETs," *Solid-State Electronics* **143**, 49-55, 2018.

[11] S. Jin, T.-W. Tang, and M. V. Fischetti, "Simulation of Silicon Nanowire Transistors Using Boltzmann Transport Equation Under Relaxation Time Approximation," *IEEE Trans. Electron Devices* **55**(3), 727-736, 2008.

[12] A. Svizhenko, M. P. Anantram, T. R. Govindan, *et al.*, "Two-dimensional quantum mechanical modeling of nanotransistors," *Journal of Applied Physics* **91**, 2343, 2002.

[13] M. Luisier and G. Klimeck, "Atomistic full-band simulations of silicon nanowire transistors Effects of electron-phonon scattering," *Phys Rev B* **80**, 155430, 2009.

[14] S. Berrada, H. Carrillo-Nuñez, J. Lee, *et al.*, "Nano-Electronic Simulation Software (NESS): a flexible nano-device simulation platform," *Journal of Computational Electronics* **19**, 1031–1046, 2020.

[15] J. Lee, O. Badami, H. Carrillo-Nuñez, *et al.*, "Variability Predictions for the Next Technology Generations of n-type SixGe1-x Nanowire MOSFETs," *Micromachines* **9**(12), 643, 2018.

[16] J. Lee, S. Berrada, H. Carrillo-Nuñez, *et al.*, "The Impact of Dopant Diffusion on Random Dopant Fluctuation in Si Nanowire FETs: A Quantum Transport Study," *2018 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 280-283, 2018.

[17] T. Dutta, C. Medina-Bailon, H. Carrillo-Nuñez, *et al.*, "Schrödinger Equation Based Quantum Corrections in Drift-Diffusion: A Multiscale Approach," *IEEE Nanotechnology Materials and Devices Conference (NMDC)*, 2019.

[18] C. Medina-Bailon, O. Badami, H. Carrillo-Nuñez, *et al.*, "Enhanced Capabilities of the Nano-Electronic Simulation Software (NESS)," *2020 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, September 2020.

[19] T. Sadi, C. Medina-Bailon, M. Nedjalkov, *et al.*, "Simulation of the Impact of Ionized Impurity Scattering on the Total Mobility in Si Nanowire Transistors," *Materials* **12**(1), 124, 2019.

[20] O. Badami, S. Berrada, H. Carrillo-Nuñez, *et al.*, "Surface Roughness Scattering in NEGF using self-energy formulation," *2019 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2019.

[21] O. Badami, T. Sadi, F. Adamu-Lema, *et al.*, "A Kinetic Monte Carlo study of retention time in a POM molecule-based flash memory," *IEEE Transactions on Nanotechnology*, Early Access, 2020.

[22] H. Carrillo-Nuñez, C. Medina-Bailon, V.P. Georgiev, *et al.*, "Full-band quantum transport simulation in presence

of hole-phonon interactions using a mode-space k·p approach," *Nanotechnology Journal*, Early Access, 2020.

[23] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFETs: A 3D “atomistic” simulation study," *IEEE Trans. Electron Devices* **45**(12), 2505–2513, 1998.

[24] S. Kaya, A.R. Brown, A. Asenov, *et al.*, "Analysis of Statistical Fluctuations due to Line Edge Roughness in sub-0.1μm MOSFETs," *Simulation of Semiconductor Processes and Devices*, Springer, 2001.

[25] A. Brown, J. Watling and A. Asenov, "Intrinsic parameter fluctuations due to random grain orientations in high-k gate stacks," *J. Comput. Electron.* **5**(4), 333-336, 2006

[26] D. L. Scharfetter and D. L. Gummel, "Large signal analysis of a Silicon Read diode oscillator," *IEEE Trans. Electron Devices*, **16**, 64-77, 1969.

[27] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in Arsenic-, Phosphorus-, and Boron-doped Silicon," *IEEE Trans. Electron Devices* **30**(7), 764–769, 1983.

[28] K. Yamaguchi, "Field-dependent mobility model for two-dimensional numerical analysis of MOSFET’s," *IEEE Trans. Electron Devices* **26**(7), 1068–1074, 1979.

[29] D. Caughey and R. Thomas, "Carrier mobilities in silicon empirically related to doping and field," *Proceedings of the IEEE* **55**(12), 2192–2193, 1967.

[30] G. Fiori and G. Iannaccone, "Three-dimensional simulation of one-dimensional transport in silicon nanowire transistors," *IEEE Transactions on Nanotechnology* **6**(5), 524–529, 2007.

[31] G.A. Kathawala and U. Ravaioli, "3-D monte-carlo simulations of FinFETs," *IEEE International Electron Devices Meeting (IEDM)*, 29–2, 2003.

[32] A. R. Brown, J. R. Watling, G. Roy, *et al.*, "Use of density gradient quantum corrections in the simulation of statistical variability in MOSFETs," *Journal of Computational Electronics* **9**(3-4), 187–196, 2010.

[33] C. Medina-Bailon, T. Sadi, M. Nedjalkov, *et al.*, "Study of the 1D Scattering Mechanisms' Impact on the Mobility in Si Nanowire Transistors," *2018 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, 2018.

[34] C. Medina-Bailon, T. Sadi, M. Nedjalkov, *et al.*, "Impact of the Effective Mass on the Mobility in Si Nanowire Transistors," *2018 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 297-300, 2018.

[35] L. V. Keldysh, "Diagram Technique for Nonequilibrium Processes," *Sov. Phys. JETP* **20**(4), 1018, 1965.

[36] G. D. Mahan, "Many-Particle Physics, Physics of Solids and Liquids," *Springer*, 2000.

[37] G. Stefanucci and R. Van Leeuwen, "Nonequilibrium Many-Body Theory of Quantum Systems: A Modern Introduction, " *Cambridge University Press*, 2013.

[38] C. Medina-Bailon, H. Carrillo-Nuñez, J. Lee, *et al.*, "Quantum Enhancement of a S/D Tunneling Model in a 2D MS-EMC Nanodevice Simulator: NEGF Comparison and Impact of Effective Mass Variation," *Micromachines* **11**(2), 204, 2020.

[39] H. Carrillo-Nuñez, J. Lee, S. Berrada, *et al.*, "Random Dopant-Induced Variability in Si-InAs Nanowire Tunnel FETs: A Quantum Transport Simulation Study", *IEEE Electron Device Lett*. **39**(9), 1473-1476, 2018.

[40] M. P. Anantram, "Modeling of nanoscale devices, " *Proc. IEEE* 96, 1511–1550, 2008.

[41] M. Luisier, A. Schenk, W. Fichtner, "Quantum transport in two- and three-dimensional nanoscale transistors: coupled mode effects in the nonequilibrium Green’s function formalism, " *J. Appl. Phys.* **100**(4), 043713, 2006.

[42] M. Aldegunde, A. Martinez, A. Asenov, "Non-equilibrium Green’s function analysis of cross section and channel length dependence of phonon scattering and its impact on the performance of Si nanowire field effect transistors, " *J. Appl. Phys.* **110**(9), 094518, 2011.

[43] M. Bescond, C. Li, H. Mera, et al., "Modeling of phonon scattering in n-type nanowire transistors using one-shot analytic continuation technique, " *J. Appl. Phys.* **114**(15), 153712, 2013.

[44] C. Jacoboni, L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys.* **55**, 645–705, 1983.

[45] A. Esposito, F. Martin, A. Schenk, "Quantum transport including nonparabolicity and phonon scattering: application to silicon nanowires," *J. Comput. Electron.* **8**(3), 336, 2009.

Photography & Biography

Cristina Medina-Bailon received the Ph.D. in Electronics from the Univ. of Granada, Spain, in 2017. Her current research interests focus on the Monte Carlo description of tunneling phenomena and on the implementation of classical and semi-classical approaches. In July 2017, she joined the Device Modelling Group at the University of Glasgow as Post-doctoral fellow and she is the software coordinator of NESS since March 2019.

Tapas Dutta received the Ph.D. degree in nanoelectronics and nanotechnology from the Grenoble INP, France in 2014. He was with IIT Kanpur, India as a postdoctoral researcher during 2014-17. Since September 2017, he has been with the Device Modeling Group at the University of Glasgow, where he has been a co-developer of NESS. His research interests are TCAD software development, and compact modeling of emerging electronic devices.

Fikru Adamu-Lema received the Ph.D. degree in electronics engineering from the University of Glasgow in 2006. He then joined Visual Numerics International, Ltd., (now a part of Rogue Wave software) as a Technical Engineer for IMSL numerical algorithms. He is currently an RA with the Device Modelling Group at the University of Glasgow and Semiwise Ltd. His current research interests include the development of reliability models, and statistical simulation study of nanoscale MOSFETs and SRAM circuit simulations.

Ali Rezaei received his Ph.D. (Dr. rer. nat.) in Condensed Matter Physics at the University of Konstanz, Konstanz, Germany, in 2019. In August 2020, he joined the Device Modelling Group, School of Engineering at the University of Glasgow as a Postdoctoral Research Associate to focus primarily on the further development and expanding the functionality of the non-equilibrium Green’s function (NEGF) quantum transport module of NESS.

Daniel Nagy received the M.Res. degree in nanoscience to nanotechnology and the Ph.D. degree in electronic and electrical engineering from Swansea University, Swansea, U.K., in 2013 and 2016, respectively. He held a Post-Doctoral position between 2016 and 2019 at the CITIUS, University of Santiago de Compostela, Santiago de Compostela, Spain. He joined the Device Modelling Group, School of Engineering, University of Glasgow, where he is a Research Associate since 2020 April.

Vihar P. Georgiev received his Ph.D. degree from the University of Oxford, Oxford, U.K., in 2011. In 2011, he joined the Device Modelling Group, School of Engineering, University of Glasgow, where he was a Research Associate until 2015 and currently he is a Senior Lecturer in electronics and nanoscale engineering. He is also UKRI EPSRC innovation fellow.

Asen Asenov (M’96–SM’05–F’11) received the Ph.D. degree in solid-state physics from the Bulgarian Academy of Sciences, Sofia, Bulgaria, in 1989. He was a Chief Executive Officer with Gold Standard Simulations, Ltd., Glasgow, U.K. Currently, he is a James Watt Professor of Electrical Engineering with the University of Glasgow.

J. Microelectron. Manuf. **3**, 20030405 (2020)
doi: 10.33079/jomm.20030405

First-principles Simulations of Tunneling FETs Based on van der Waals $MoTe_2/SnS_2$ Heterojunctions with Gate-to-drain Overlap Design

Kun Luo[1,2], Kui Gong[2], Jiangchai Chen[2], Shengli Zhang[3,*], Yongliang Li[1],
Huaxiang Yin[1], Zhenhua Wu[1,2, **]

[1] Key Laboratory of Microelectronics Device and Integrated Technology, Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China,
[2] IMECAS-MU-HZW joint computing laboratory of Integrated Circuits, Beijing 100029, China
[3] MIIT Key Laboratory of Advanced Display Materials and Devices, School of Materials Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

Abstract: The electronic properties and transport properties of $MoTe_2/SnS_2$ heterostructure Tunneling FETs are investigated by the density functional theory coupled with non-equilibrium Green's function method. Two dimensional (2D) monolayer $MoTe_2$ and SnS_2 are combined to a vertical van der Waals heterojunction. A small staggered band gap is formed in the overlap region, while larger gaps remain in the underlap source and drain regions of monolayer $MoTe_2$ and SnS_2 respectively. Such a type-II heterojunction is favorable for tunneling FET. Furthermore, we suggest short stack length and large gate-to-drain overlap to enhance the on-state current suppress the leakage current respectively. The numerical results show that at a low drain to source voltage V_{ds} = 0.05V, On/Off current ratio can reach 10^8 and the On-state currents is over 20 µA/µm for n-type devices. Our results present that van der Waals heterostructure TFETs can be potential candidate as next generation ultra-steep subthreshold and low-power electronic applications.

Keywords: 2D materials heterojunction, tunnel-FET, gate-to-drain overlap, DFT-NEGF.

1. Introduction

The downscaling of field-effect-transistors (FETs) to sub-5nm and more advanced technical node is following the Moore's law and approaching their physical limitations with traditional silicon FETs. Recently, the discovery of two-dimensional materials in 2005 [1], has opened a brand-new concept to semiconductor engineers who are seeking new materials for replacing the silicon and improving the performance of semiconductor device. Two-dimensional (2D) material-based semiconductor has been acknowledged as a promising option for the next-generation electronics because of their uniform atomic thickness, smooth surface and excellent gate electrostatic controlling ability. With the development of the significant advances in nanotechnology, in the past few years, 2D material field effect transistors (FETs) have drawn a lot of attentions with several 2D materials, such 2D MoS_2 [2-4], 2D InSe [5, 6], black phosphorus (BP) [7-9], 2D Bi_2O_2Se [10] and so on [11-14].

The potential of these materials has not been thoroughly investigated, and the development of manufacturing atomically thin van der Walls heterostructures gives rise to new opportunities [11, 14]. More and more experimental works have focused on the properties of plane heterojunction and stacked heterojunction [3]. According to the recently researches, high quality 2D SnS_2-based FETs have been measured and their ultrahigh on/off current ratio can reach to 10^8, which is higher than that of BP and other 2D materials FETs [15-16]. As for 2D $MoTe_2$ material, it has been fabricated all-2D-based FETs which also can reach quite high mobility (over 20 $cm^2V^{-1}s^{-1}$) and on/off current ratio about 10^5 [17]. However, 2D materials FETs need to satisfy high speed and low energy dissipation applications, which means a lot of challenges exist [18]. As an alternative application, band-to-band tunneling FETs combine with stacked 2D heterojunction can be potential candidates. Importantly, TFETs can make a breakthrough in subthreshold slope (SS) reduced below 60 mV/dec and have a quite low OFF-state current [19]. In ultra-thin vertical heterojunctions, the tunneling distance is reduced to the minimum, which affords the possibility to achieve higher ON-state

[*] Address all correspondence to Shengli Zhang, E-mail: zhangslvip@njust.edu.cn
[**] Address all correspondence to Zhenhua Wu, E-mail: wuzhenhua@ime.ac.cn

current. Furthermore, 2D TFETs can effectively control the leakage of direct source-to-drain tunneling and do not have an influence on band-to-band tunneling because of the staggered band alignment when two layers are stacked together [20-22]. Finally, it is expected that a higher on/off current ratio and lower SS will be achieved while the tunneling occurs between two different monolayer 2D materials.

In this work, we investigate a stacked heterojunction tunneling FET based on van der Waals $MoTe_2/SnS_2$ heterojunctions (see Figure 1) with gate-to-drain overlap. $MoTe_2$ and SnS_2 are two semiconductors with relatively larger band gaps and their stacked structure has the staggered band alignment which is desired to achieve high on-state tunneling current with acceptable leakages [23-24]. The two materials have high carrier mobility, i.e., hole mobility of MoTe2 is about 200 $cm^2V^{-1}s^{-1}$ [25] and the electron mobility of SnS2 can reach about 1398 $cm^2V^{-1}s^{-1}$ [26]. The type-II heterojunction with a small staggered band gap is formed for tunneling transistors [20]. Furthermore, short stack length and large gate-to-drain overlap (see Figure 2) are proposed to enhance the on-state current suppress the leakage current respectively. We employ the density function theory (DFT) method to study the basic electronic properties of monolayer $MoTe_2$ and SnS_2. Then, the transport properties of the double gate stacked structure are calculated by Non-Equilibrium Green's Function (NEGF) method. The merits of the proposed TFET, including local density of state (LDOS), on-state current and SS, are compared with monolayer $MoTe_2$ n-TFET. The device performance of the $MoTe_2$-SnS_2 TFETs presents the great potential for future semiconductor applications.

2. Simulations Methods

Most of previous studies utilize Tight-binding Non-Equilibrium Green's Function (TB-NEGF) method to predict the device performance with new materials and operation mechanisms. It is a good compromise between the computational costs and the coverage of quantum transport feature. For example, one typical TB Hamiltonian employs Slater-Koster (SK) parameters by fitting the electronic structure from DFT method [27]. The transport properties are calculated utilizing the fully quantum mechanical NEGF formalism. Note that DFT includes exchange correlation potentials as well as external potentials, to generate the accurate energy band. However, based on SK parameters by fitting the band of DFT, TB only considers the external potential to self-consistently solve the potential field. The calibration and setup of TB parameter library for new materials can be tedious and tricky. In this work, the calibration-free DFT-NEGF method is used to investigate the tunneling FETs based on van der Waals $MoTe_2/SnS_2$ heterojunctions, i.e., using the DFT to calculate the Hamiltonian and electrostatic properties of the device; using NEGF to determine non-equilibrium statistics for constructing density matrix; using real Space numerical methods to calculate transport properties and the boundary conditions for open device structures [28]. High precision can be achieved by using DFT-NEGF, but at expense of computational issues in speed and memory limits.

At present, the mainstream DFT-NEGF programs are able to simulate 5000 atomic-scale structures or devices effectively, but larger-scale computational simulations still have difficulties to overcome. If the number of atoms is further increased, there will be insufficient memory. The scale of parallel processes is another limitation. The parallel computing efficiency is poor as the employed CPU cores are increased. This is mainly due to the inefficient use of computing resources. We develop the DFT-NEGF calculation method for this specific application of $MoTe_2/SnS_2$ TFET in the following aspects, 1) the matrix distributed calculation mode is introduced; 2) optimize the Poisson equation and Green's function solution algorithm under specific boundary conditions; 3) at the same time, optimize the linear combination of atomic orbitals (LCAO) basis [28-29] set of each element involved in the tunneling transistor to reduce the matrix dimension without reducing the calculation accuracy and improve the calculation ability. The details are beyond the scope of this paper and will be reported elsewhere. The calculations in this work are based on the Nanodcal packages with the aforementioned updates [28].

3. Results and Discussions

In order to precisely calculate the electronic states properties of the $MoTe_2$-SnS_2 heterojunction model, we employ the DFT based ab-initio package Nanodcal. The generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof (PBE) is applied for the exchange-correlation interactions, which can exactly present band gap values in good agreement with experiment results for monolayer 2D materials. The energy cutoff is 500 eV and the

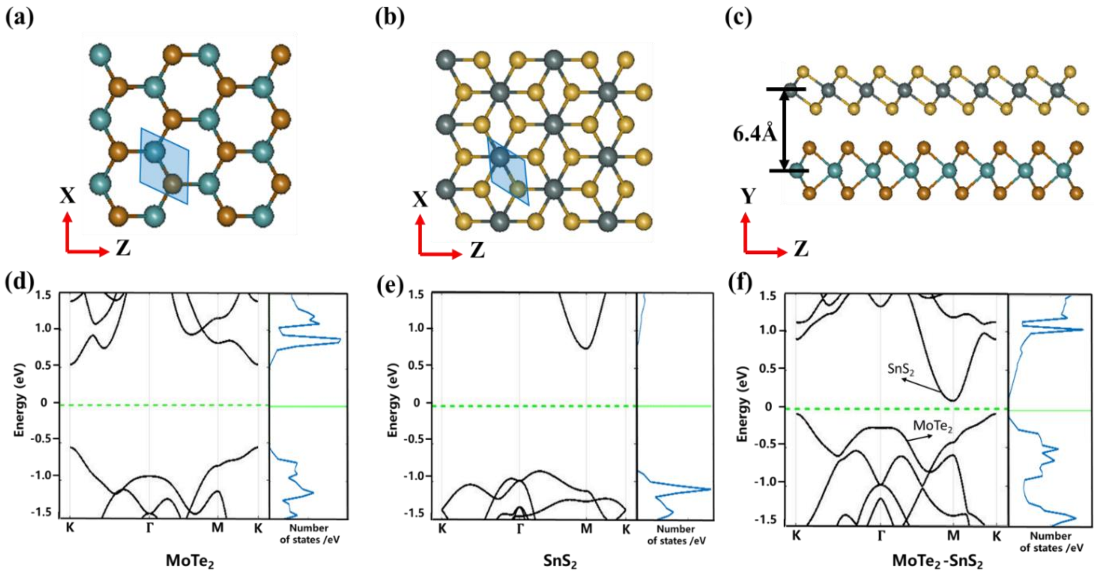


Figure 1. (a-c) basic structure of $MoTe_2$, SnS_2 and heterostructure, the permittivity cell is covered by the shadow area contained three atoms; (d-f) Band structure and density of states of intrinsic $MoTe_2$, SnS_2 and $MoTe_2$-SnS_2 heterostructure.

Monkhorst-Pack k points are set as 9 x 9 x 1 without spin-orbit coupling. The convergence criteria for energy and force are 10^{-4} eV and 10^{-3} eV/ Å. The relaxed monolayer $MoTe_2$ and SnS_2 are shown in Figure 1 (a) and (b) with the lattice constants being 3.56 Å and 3.70 Å respectively. And the heterostructure is built after applying strain to both two materials so as to obtain the same lattice parameter a_0 = 3.625 Å as shown in Figure 1 (c). To study the basic properties of monolayer $MoTe_2$ and SnS_2, the band structure of two materials is calculated along the high-symmetry path (K-Γ-M-K) in Brillouin zones. As shown in Figure 1 (d)-(f), the band structure of intrinsic monolayer $MoTe_2$ has a 1.10 eV direct band gap at K point, like the other traditional 2D semiconductor materials. And monolayer SnS_2 has an indirect gap of about 1.61 eV that is an applicable value as the channel material of MOSFETs. Combined two materials, it formed a system that is a type-II heterojunction with a 0.29 eV indirect band gap which is larger than the band gap of another similar combination of 2D material stack, *i.e.*, WTe2-MoS2 (0.16 eV) [23]. Compared all three band structures, it is obvious that the valence band maximum (VBM) is contributed by $MoTe_2$ at K and the conduction band minimum (CBM) is contributed by SnS_2 at M. Therefore, if the transport axis is along the M-K direction, the momentum is conserved in the periodic direction and the tunneling process can be formed along the transport direction due to the variation of electrostatic potential.

The device band edges schematics demonstrate the mechanism of the $MoTe_2$-SnS_2 TFETs as shown in Figure 2 (a). The Type II band alignment can effectively keep the tunneling window of channel. Note that the interaction of the stacking edge has a dramatically deviation of the band structure, as compared with monolayer or heterostructure, which dominate the tunneling on current of the TFET. Longer heterostructure length can hardly enhance the tunneling on current since the bands are rather flat in the middle region, but leads to larger channel resistance. On the other hand, very short heterostructure length also leads to TFEF performance degradation due to the direct source to drain tunneling.

To obtain multi-objective optimization for on-state current and off-state current trade-offs, the schematics of TFETs device with gate-to-drain overlap design is presented in Figure 2(b). The distance between $MoTe_2$ and SnS_2 layers is 6.4Å, which is optimized by optB86 exchange correlation functional method. To avoid the influence of mismatch, monolayer $MoTe_2$ is applied 2.1% tensile strain and 2.0% compressive strain is for SnS_2, which can keep lowest mismatch. For the whole TFET device structure, the total number of atoms exceeds 360. As shown in Figure 2 (b), the out-of-plane vacuum separation of the device is fixed as 2 nm, which is equal to the distance between the top and bottom gates. Spin-orbit coupling is excluded. In our work, we investigate the influence of EOT

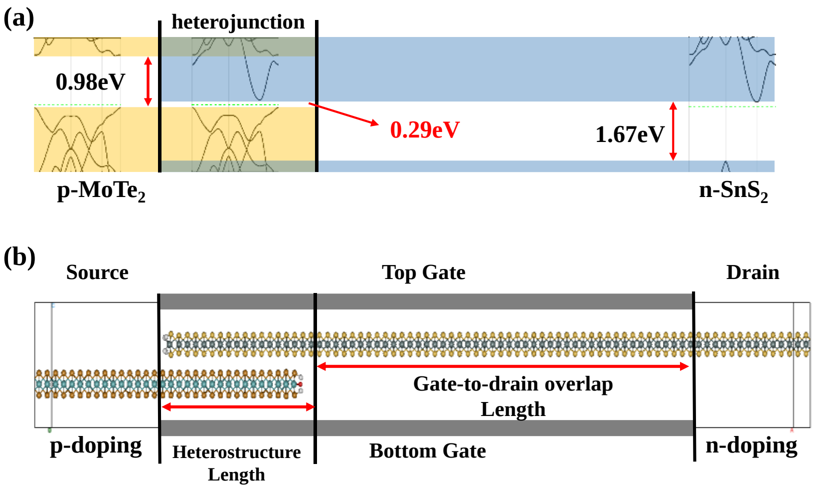


Figure 2. (a) Band alignment schematic in the device along the transport direction with flat band condition; (b) Schematic of the double-gate MoTe2-SnS2 heterostructure TFETs with gate to drain overlap.

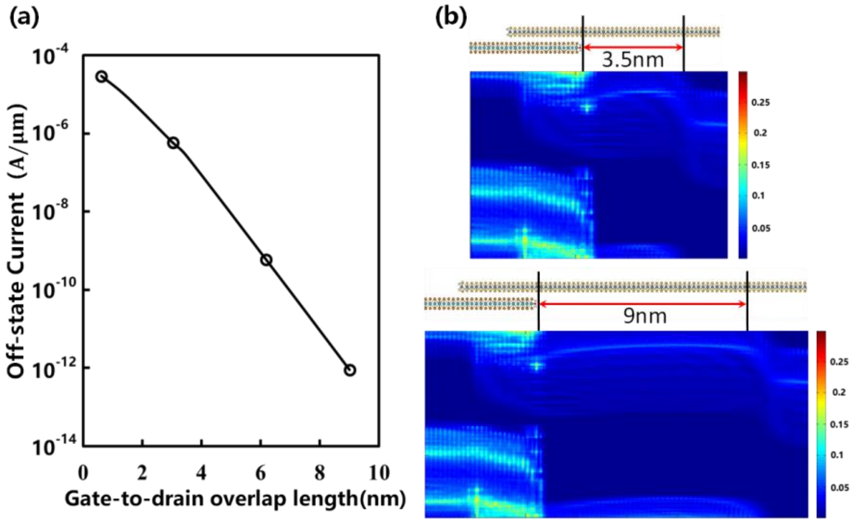


Figure 3. (a) The variation of off-state current with different length of gate-to-drain overlap between heterostructure to drain. With the increasing of length, the leakage of drain can be effectively suppressed at the same supply voltage; (b) LDOS of the TFETs with different gate-to-drain overlap length at V_{gs} = 0V.

variation for the performance of TFETs device. The default effective oxide thickness (EOT) is set to 0.5 nm with effective κ=3.9. For the gate voltage, the bias is only applied to the top gate and the bottom is set as ground. In the case of n-type device, the source side is doped to p-type and the drain side is doped to n-type. Both the source and drain doping concentration reach 10^{13} cm^{-2}. And the intrinsic materials are employed for the channel because the device performance is insensitive to the doping concentrations. The supply voltage is set as V_{ds} = 0.05 V in all the following simulation. In this condition, the self-consistent electrostatic and transport calculation for each gate bias point spends about 18 hours of wall time by using 144 CPU cores.

Firstly, we investigate the impact of length variation from heterostructure to drain (gate-to-drain overlap) on the device properties. The heterojunction length is fixed to 3 nm. As shown in Figure 3, the leakage can be effectively reduced with increasing the length of overlap. At V_{gs} = 0 V, the Off-state current of 4 nm overlap is as large as 10^{-7}A/μm and it can be reduced to 10^{-13}A/μm with 9 nm overlap condition. It indicates that electrons in the VB of $MoTe_2$ have a high probability of tunneling into the CB of SnS_2 without gate voltage at short overlap region. The gate-to-drain overlap design gives rise to good optimization for Ion and Ioff trade-offs as compared to normal TFET in previous studies [20] (see Table. 1). The 9 nm gate-to-drain overlap

Table 1. Heterojunction length and corresponding device key merits.

Heterojunction Length (nm)	Ioff (A/μm)	Ion (A/μm)	Ion/Ioff	SS (mV/dec)
3 (this work)	2.05E-12	5.77E-04	10^8	37
6 (this work)	2.38E-14	1.46E-04	10^{10}	42
20 [Ref 20]	1E-12	7.5E-5	10^7	<60

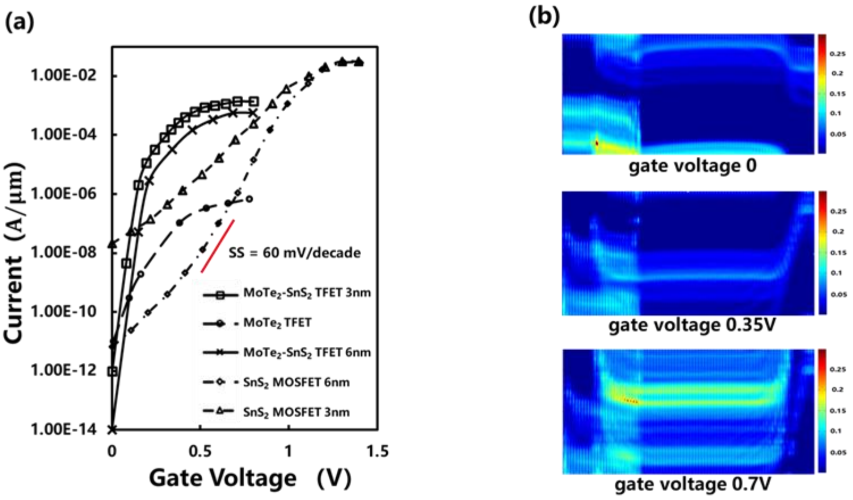


Figure 4. (a) I_d -V_{gs} transfer properties. The solid lines represent the transport characteristics of heterojunction with different heterojunction length region and the dashed lines demonstrate transport properties of single-layer $MoTe_2$ TFET; (b) LDOS of the TFETs with different gate voltages.

structure is selected as the basic TFET structure to calculate following transport characteristic. In addition, as increasing EOT from 0.5 nm to 1 nm, the performance of TFET device present the degradation tendency, e.g., the subthreshold slope drops to 48 mV/dec and the Ion/Ioff ratio decreases about an order of magnitude. To achieve optimal performance, thin EOT of 0.5 nm is selected in the following calculations.

Then, the transport properties of n-type device are simulated by the DFT-NEGF method. The I_d-V_{gs} curve of the $MoTe_2$-SnS_2 TFETs is shown in Figure 4 (a). It is obviously that the sub 60 mV/decade subthreshold swing is obtained as about 37 mV/decade. By fixing the Off-state current of the device to 10^{-6} μA/μ*m*, the current can achieve about 20 μA/μ*m*. For benchmark, the transfer characteristics of a single-layer $MoTe_2$ TFET is also simulated. The SS is not notably below the limitation of MOSFET and the On/Off current ratio only reach to about 10^5 due to the short channel length. The mechanism of n-type TFETs is demonstrated in Figure 4 (b), which presents the LDOS of the TFETs with the different gate voltage. At V_{gs} = 0V, tunneling path does not exist because the VBM of $MoTe_2$ is located below the CBM of SnS_2 in the channel region, which also proves that the buffer layers of both two side is long enough to keep the minimal impact of the leakage. With the increasing of gate voltages, the CBM of SnS_2 is dropped down faster than the VBM of $MoTe_2$, on account of the effectively modulation of SnS_2. At V_{gs} = 0.35V, electrons in the VB of $MoTe_2$ are gradually enter into the CB of SnS_2 at the center of the channel. They can tunnel from the source of $MoTe_2$ cell into the drain because the CBM is getting lower in the monolayer than in the heterostructure. In the On-States, the bands of $MoTe_2$ and SnS_2 are totally changed in the overlap region when the gate voltage is over 0.4 V. Electrons can freely cross through the whole stacking area at high gate voltage. This indicates that the $MoTe_2$-SnS_2 TFETs can have better performance compared with single-layer 2D materials MOSFETs.

Based on this result, we further investigate the influence of increasing the heterojunction region length from 3 nm to 6 nm. On one hand, the length of the stacking region cannot markedly enhance the on-state current. On the other hand, it leads to a better control of the off-state current as shown in Figure 5. As the data presented in Table 1, the on-state current of short heterostructure length TFET is

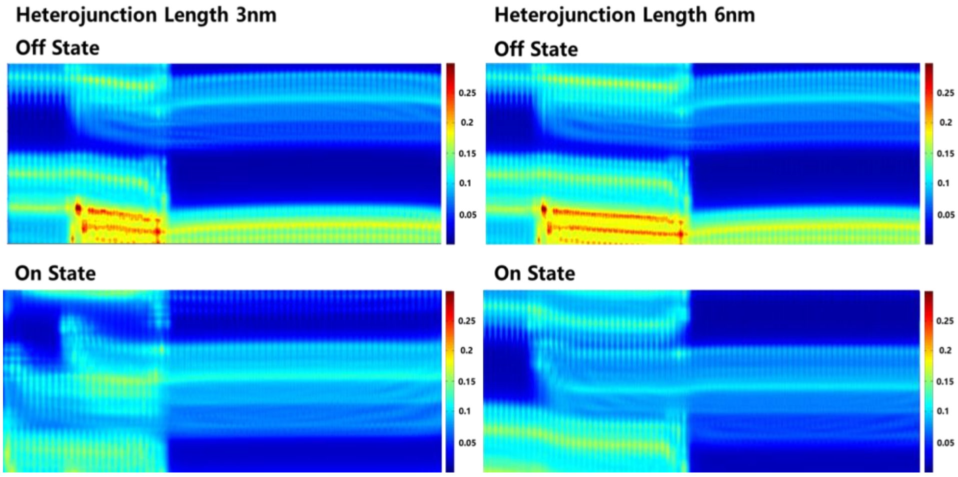


Figure 5. LDOS of TFETs with different heterojunction length at On/Off states. At V_{gs} = 0V, the device of 6nm presents more effective.

similar to the longer heterojunction length. The band-to-band tunneling and the direct source-to-drain tunneling are essential in the tunneling processes. The longer heterojunction length device can be effectively suppressed the leakage arising from the direct source-to-drain tunneling, but does not notably affect the on-state current due to the band-to-band tunneling.

4. Conclusion

In this work, we investigate the electronic properties of a $MoTe_2$-SnS_2 heterostructure by DFT, which gives rise to a small staggered gap in the stack overlap region and large gap in the source drain. Based on this heterostructure, a double-gate n-type TFET with gate-to-drain overlap has been designed and calculated by DFT-NEGF. At a low supply voltage V_{ds} = 0.05V, On/Off current ratio reaches to 10^8 and the subthreshold swing is well below the thermal limitation of traditional Silicon MOSFET. It is reasonable that 2D van der Waals heterostructures have a great potential in next generation of ultra-steep subthreshold and low-power applications.

Acknowledgments

This research financially supported by the Training Program of the Major Research Plan of the National Natural Science Foundation of China (61774168, 91964103) and the MOST (2016YFA0202300).

References

[1] K. Novoselov, D. Jiang, F. Schedin, et al. “Two-dimensional atomic crystals,” PNAS, 102, 10451-10453 (2005).
[2] B. Radisavljevic, A. Radenovic, J. Brivio, et al. “Single-Layer MoS2 Transistors,” Nat. Nanotechnol, 6, 147−150 (2011).
[3] X. Hong, J. Kim, Y. Zhang et al. “Ultrafast charge transfer in atomically thin MoS2/WS2 heterostructures,” Nature nanotechnology, 9, 682-686 (2014).
[4] Y. Pan, Z. Wu, H. Yin, et al. “Near-ideal subthreshold swing MoS2 back-gate transistors with an optimized ultrathin HfO2 dielectric layer”. Nanotechnology, 30, 9, 095202 (2019).
[5] P. Ho, Y. Chang, Y. Chu, et al. “High-Mobility InSe Transistors: The Role of Surface Oxides,” ACS Nano., 11:7362-7370, (2017).
[6] K. Luo, W. Yang, Y. Pan. et al. “Ab-Initio Simulations of Monolayer InSe and MoS2 Strain Effect: From Electron Mobility to Photoelectric Effect,” Journal of Elec Materi, 49, 559–565 (2020).
[7] R. Zhang, Z. Wu, X. Li, et al. “Fano resonances in bilayer phosphorene nanoring”. Nanotechnology, 29, 21, 215202 (2018).
[8] X. Li, K. Luo, Z. Wu, et al. “Tuning the electrical and optical anisotropy of a monolayer black phosphorus magnetic superlattice”. Nanotechnology, 29, 17 174001 (2018).
[9] N. Haratipour, S. Namgung, S. Oh, et al. “Fundamental Limits on the Subthreshold Slope in Schottky Source/Drain Black Phosphorus Field-Effect Transistors,” ACS Nano, 10, 3, 3791–3800 (2016).
[10] P. Luo, F. Wang, K. Liu, et al. “PbSe Quantum Dots Sensitized High-Mobility Bi2O2Se Nanosheets for High-Performance and Broadband Photodetection Beyond 2 mum.” ACS Nano, 13, 9028-9037 (2019).
[11] A. K. Geim and I. V. Grigorieva, “Van der Waals heterostructures,” Nature, 499, 419–425 (2013).
[12] Z. Wu. “Electronic fiber in graphene”. Appl. Phys. Lett., 98,082117 (2011).
[13] Z. Wu, F. Zhai, K. Chang, et al. “Valley-Dependent Brewster Angles and Goos-Ha¨nchen Effect in Strained Graphene”. Phys. Rev. Lett., 106, 176802 (2011).

[14] T. Brumme, M. Calandra, F. Mauri, “First-principles theory of field-effect doping in transition-metal dichalcogenides: structural properties, electronic structure, Hall coefficient, and electrical conductivity,” Phys. Rev. B, 91, 155436 (2015).

[15] S. Wei, C. Ge, L. Zhou, et al., “Performance Improvement of Multilayered SnS2 Field Effect Transistors through Synergistic Effect of Vacancy Repairing and Electron Doping Introduced by EDTA,” ACS Appl. Electron. Mater. 1, 11, 2380–2388 (2019).

[16] D. Chu, S. Pak, E. Kim, “Locally Gated SnS2/hBN Thin Film Transistors with a Broadband Photoresponse,” Sci Rep, 8, 10585 (2018).

[17] K. Choi, Y. Lee, J. Kim, et al. “Non‐Lithographic Fabrication of All‐2D α‐MoTe2 Dual Gate Transistors,” Adv. Funct. Mater. 26: 3146-3153 (2016).

[18] C. Klinkert, Á. Szabó, M. Luisier*, et al. “2-D Materials for Ultrascaled Field-Effect Transistors: One Hundred Candidates under the Ab Initio Microscope,” ACS Nano, 14, 7, 8605–8615 (2020).

[19] Q. Zhang, G. Iannaccone and G. Fiori, "Two-Dimensional Tunnel Transistors Based on Bi2Se3 Thin Film," in IEEE Electron Device Letters. 35, 1, 129-131 (2014).

[20] Á. Szabó, S. J. Koester and M. Luisier, "Ab-Initio Simulation of van der Waals MoTe2–SnS2 Heterotunneling FETs for Low-Power Electronics," in IEEE Electron Device Letters. 36, 5, 514-516 (2015).

[21] K. Lam, X. Cao, and J. Guo, “Device performance of heterojunction tunneling field-effect transistors based on transition metal dichalcogenide monolayer,” IEEE Electron Device Lett. 34, 10, 1331–1333 (2013).

[22] L. Xu, P. Zhang, Z. Hu, et al. “Large‐Scale Growth and Field‐Effect Transistors Electrical Engineering of Atomic‐Layer SnS2,” Small, 15, 1904116 (2019).

[23] J. Cao, M. Pala, D. Esseni, et al, “Operation and Design of van der Waals Tunnel Transistors: A 3-D Quantum Transport Study”, IEEE Transactions on Electron Devices, 63, 11, 4388-4394 (2016).

[24] Á. Szabó, C. Klinkert, D. Campi, et al, "Ab Initio Simulation of Band-to-Band Tunneling FETs With Single- and Few-Layer 2-D Materials as Channels," IEEE Transactions on Electron Devices, 65, 10, 4180-4187 (2018).

[25] S. Mir, V. Yadav, and J. Singh. “Recent Advances in the Carrier Mobility of Two-Dimensional Materials: A Theoretical Perspective”, ACS Omega, 5, 24, 14203-14211 (2020).

[26] A. Shafique, A. Samad and Y. Shin. “Ultra low lattice thermal conductivity and high carrier mobility of monolayer SnS_2 and $SnSe_2$: a first principles study”, Phys. Chem. Chem. Phys., 19, 20677-20683 (2017).

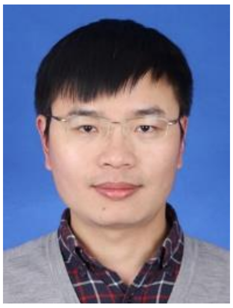
[27] A. D. Becke, “Density-functional exchange-energy approximation with correct asymptotic behavior,” Phys. Rev. A, 38, 3098 (1988).

[28] J. Taylor, H. Guo, and J. Wang, “Ab initio modeling of open systems: Charge transfer, electron conduction, and molecular switching of a C60 device,” Phys. Rev. B, 63, 121104 (2001).

[29] J.E. Huheey, E. Keiter, R. Keiter,et al. “Inorganic Chemistry: Principles of Structure and Reactivity”, 4th ed., Dorling Kindersley Pvt Ltd (2008).

Photography & Biography

Kun Luo received the M.S. degree of physics from University of Durham, Durham, United Kingdom and the B.S. degree in physics from Shandong University, Jinan, China, in 2015,. In 2017, he became a research intern in Institute of Microelectronics of the Chinese Academy of Sciences. His current research interest includes first principle and TCAD model.



Kui Gong received the Ph.D. degree in Material Science from the University of Science & Technology Beijing，Beijing, China, in 2015. Meanwhile, as a Joint Ph.D. student research in Condensed Matter Physics department of McGill University, Montreal, Canada, during the period of 2012-2015. He jointed Tech Department, Hongzhiwei Technology (Shanghai) CO., LTD, China, in 2016. Now, he is the manager of Application Technology Department of Hongzhiwei Tech. His current research interests include Quantum transport simulation, TCAD, Density functional theory.

Jiangchai Chen received the Ph.D degree in Theory Physics from Institute of Physics, Chinese Academy of Sciences, Beijing, China, in 2012. Then he did one-year post-doctor research in the Department of Physics, the University of Hong Kong. In 2015, he joined Technology Department (now the R&D Center), Hongzhiwei Technology (Shanghai) Co., Ltd, Shanghai, China. He is currently responsible for the development of a quantum transport software, Nanoskim.

Shengli Zhang received his PhD degree from Beijing University of Chemical Technology in 2013. He then joined the Key Laboratory of Advanced Display Materials and Devices, Nanjing University of Science and Technology, where he is a Professor in the Department of Materials Science and Engineering. His research interests focus on electronic devices and applications based on 2D materials.

Yongliang Li received the Ph.D. degree from the Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China, in 2011. He was a Staff Engineer with the United Microelectronics Corporation, Singapore, for process integration, from 2011 to 2017. In 2018, he joined the Institute of Microelectronics, Chinese Academy of Sciences, where he is currently a Professor of engineering with the Integrated Circuit Advanced Process Center. His current research interests include high-mobility SiGe/Ge material process integration, and novel 3-D CMOS devices.

Huaxiang Yin received the B.S. degree in semiconductor physics and devices from Tianjin University, Tianjin, China, in 1996, and the M.S. and Ph.D. degrees in microelectronics and solid-state electronics from the Chinese Academy of Sciences (CAS), Beijing, China, in 1999 and 2003, respectively. From 2003 to 2010, he was with the Samsung Advanced Institute of Technology, South Korea, as a Research Staff Member. In 2010, he joined the Institute of Microelectronics, CAS, where he is currently a Professor with the Integrated Circuit Advanced Process Center and the Key Laboratory of Microelectronics Devices and Integrated Technology. His research interests include nanoscale CMOS devices, VLSI manufacture technology, 2-D materials and devices, and Si X-ray detector.

Zhenhua Wu received the Ph.D. degree in Condensed Matter Physics from the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China, in 2011. He joined Semiconductor R&D Center, Samsung Electronics, Suwon, Korea, in 2011. In 2016, he became a professor in Chinese Academy of Science in Beijing, China. His current research interests include device physics, TCAD simulation of nanoscale transistors.

J. Microelectron. Manuf. **3**, 20030406 (2020)
doi: 10.33079/jomm.20030406

Material Modeling in Semiconductor Process Applications

Boris A. Voinov[3] [*], Patrick H. Keys[1], Stephen M. Cea[1], Ananth P. Kaushik[2], Mark A. Stettler[1]

[1] Logic Technology Development, Intel Corporation, Hillsboro OR, USA, 95124
[2] Nonvolatile Memory Solutions Group, Intel Corporation, Santa Clara, California, USA, 95054
[3] Logic Technology Development, Intel Corporation, Nizhniy Novgorod, Russian Federation, 603024

Abstract: During the past decade, significant progress has been achieved in the application of material modeling to aid technology development in semiconductor manufacturing companies such as Intel. In this paper, we review examples of applications involving a complex set of material modeling tools and methodologies and share our perspective of the future of the area. Examples are given illustrating the landscape of useful physical models and approaches along with commentary addressing tool relevance and simulation efficiency issues. While the scope of this paper precludes providing in-depth details, references to more focused publications are shared. Finally, we outline how to approach constructing a general infrastructure for supporting TCAD material modeling applications.

Keywords: TCAD, atomistic modeling, density functional theory, molecular dynamics, kinetic Monte Carlo.

1. Introduction

In the past thirty years, semiconductor modeling in industrial TCAD (Technology Computer Aided Design) has undergone incredible change. Over this period, device engineering has evolved dramatically with the introduction of the FinFET, novel materials, and strain. In addition, the concentration of fabrication facilities within just a few large companies has resulted in TCAD departments becoming more active in conducting in-house research versus relying on external sources. However, the single biggest change for TCAD is undoubtedly the scale of the problems it now tackles. In the 1990's, TCAD was concerned almost solely with simulating device performance, which meant figuring out how to control short channel effects (SCE) as the gate length shrunk by simulating various S/D and well engineering options. The modeling domain for this problem was confined to 100's of nanometers. Today, the scale of problems extends over 8 orders of magnitude. Stress engineering has moved the simulation domain beyond the device itself to including neighboring structures which also impacts the mechanical stress in the channel. Parasitic effects like latch-up and reliability phenomenon such as ion strikes require even larger scale simulations. At the upper end, calculating attributes such as die temperature, which requires simulating the heat generated from every transistor and interconnect, has extended the domain to millimeters. On the small end of the scale, features of the device such as fin width are now down to a countable number of atomic layers. As a result, TCAD must rigorously calculate quantum effects such as confinement and tunneling and also fundamental material properties, which depend not only on the novel materials employed but also on the specific number of layers. At the atomic scale, the impact of defects, which can cause changes in intrinsic strain, leakage, and resistance in semiconductors and metals, are now routinely estimated with modeling. TCAD is even tasked with simulating the properties of individual molecules such as adhesion and selectivity to help down select the reagents used in process steps such as Atomic Layer Deposition (ALD). As technology continues to scale, device and process modeling is evolving into an extended materials problem.

As a result, material modeling (MM) has become an essential part of TCAD domain along side more tradition disciplines. While scaling provides the motivation for this evolution, what has made MM possible is the tremendous progress in the development of computational methods for many-body interacting systems[1] and the incredible advances in computing power [2]. The role of MM in TCAD is twofold. On one hand, MM is used to analyze the behavior of novel structures and materials on the atomic to nanometer size scale. On

[*] Address all correspondence to Boris A. Voinov, E-mail: boris.voinov@intel.com

the other hand, researchers also resort to MM when the validity of parameters used in macroscopic simulations, which are based on continuous models, become questionable, such as when materials are employed at such a minute scale that their bulk properties no longer apply. A good example of this is understanding the heat transfer across the interface of two dissimilar materials[3]. In the case of nonmetals, where phonons are responsible for the energy transfer, the continuous heat transfer (Fourier) model fails at a length scale comparable to the average phonon mean free path, giving rise to the interface thermal resistance. From the macroscopic point of view, the temperature looks discontinuous on the interface, but by employing an MM approach such as molecular dynamics (MD) simulation[4], the energy flux can be calculated and used to extract the coefficient of interface thermal resistance, which can then be inserted into the continuous heat transfer model.

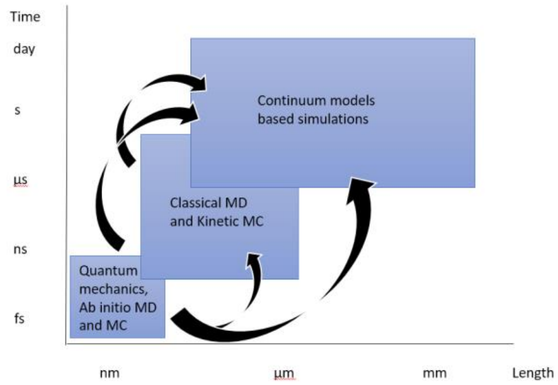

Time
day
s
μs
ns
fs
nm
μm
mm
Length
Continuum models based simulations
Classical MD and Kinetic MC
Quantum mechanics, Ab initio MD and MC

Figure 1. Time and length scales and schematics of methodologies of material modeling. Arrows show the information flow between methodologies.

Overall, the limitation of continuum models is that most assume certain constitutive relations to describe material properties, e.g. stress versus strain dependence for mechanical properties or diffusivity dependence on the temperature. These assumptions are clearly violated on the nanoscale. In contrast, when one is applying an MM approach, the only assumption is how atoms or molecules interact directly with each other. This interaction can be treated as either classically or quantum mechanically. In the classical case, the system energy is represented as a sum of contributions from a pair or many-body potentials over all atoms in the system. In the quantum treatment, the energy of electrons interacting with both ions and other electrons and the ions among themselves are calculated using the quantum theory of many-body systems. There are numerous methods for these calculations; among them, the Density Functional Theory (DFT)[5] is the most frequently used approach. It worth noting that in a TCAD context, the material subject to modeling is, as a rule, assumed to be in the solid state.

2. Applications of Material Modeling

Table 1 shows a brief list of MM applications that are relevant to semiconductor technology development and are currently in use in industrial TCAD. This list is not comprehensive and focuses primarily on applications in TCAD's traditional scope. The objective of this section is to elaborate on the content of this list.

Historically, the aim of statistical physics and condensed matter theory was the calculation of bulk properties of pure and compound homogeneous materials, where it achieved remarkable progress. To a greater degree, this progress was attributed to the fact that, for an ideal crystal, lattice electron wave functions can be relatively easy constructed as well as quantum mechanically methods to self-consistently account for non-weak many-body interactions in solids[6]. In today's DFT based computational tools, the analogous problem is choosing an appropriate electron basis function (either plane waves or atomic orbitals depending on the type of material – metal or non-metal) and a suitable form of the exchange-correlation functional[5]. These calculations don't require significant computational resources since just a single crystal cell can be used to deliver a wide variety of material properties including: the geometry of the crystal cell minimizing the total system energy and thus material density, its formation (cohesive) energy, elastic moduli as derivatives of the total energy versus the cell volume and strain, etc. A more extended theory allows the definition and quantification of the effects of elementary quantum excitations in solids – quasiparticles such as electrons, holes, and phonons[7]. Most available DFT packages[8] can calculate properties of these particles such as band structure and the phonon spectrum. With this information, one can assess transport and thermal material properties at finite temperatures such as heat capacity, thermal and electrical conductivity. This is for an ideal crystalline material, which allows the reduction of the computational domain to a single lattice cell and limits the number of atoms under consideration to just a few, explaining why the

Table 1. Material modeling applications.

Application domain	Properties of interest
Pure and compound bulk materials	Equilibrium structure, stability, equation of state, mechanical, thermophysical, heat and electric transport, electronic structure, vibration spectra
Point and extended defects	Structure, formation energy, electronic structure, optic absorption, diffusivity, fracture, plasticity
2D heterostructures, interfaces, thin films, free surfaces	Structure and defects, stability, transport, electronic structure, surface reconstruction and chemistry
Atomistic processes, etching, deposition, epitaxial growth	Selectivity, byproducts yield, effect of process conditions, microstructure formation and evolution, material damage and recrystallization
3D nanostructures	Grain structure, contacts, conductance, strain, effect of size

computational burden for this sort of calculation is modest. It should be noted that the complex quantum computations above can be bypassed if a sufficiently accurate interatomic interaction potential is known a priori which allows calculation of the total system energy. More details about this approach are described in Section 3.

The situation changes in real materials where nonuniformities, such as defects, are present.[9] Defects profoundly affect material properties on the nanoscale. In this situation, one faces a dilemma on how to construct the simulation domain, i.e. place the defect into a simulation "box" with periodic boundary conditions or insert it into a finite sample of the crystal lattice. In the first case, an artificial periodic lattice of defects will arise, which requires devising a physical way to account for their interaction energy, especially for charged defects. In the second case, one needs to extend the size of the box to ensure the results aren't sensitive to the boundary conditions. In both cases, a system with defects becomes much more computationally challenging to simulate compared to ideal crystals. Additional complications arise when we account for defect migration in realistic systems, where the positions of the lattice atoms are modulated by nearby vacancies, thus varying the potential barriers from site to site. This situation not only requires considering numerous intermediate states between the initial and final locations of the atom, but also an effective optimization technique to find the minimal energy path (MEP) associated with the transition[10]. Techniques such as the nudged elastic band (NEB) or the zero temperature string (ZTS) are available in some DFT[11] and molecular dynamics (MD)[12] packages. An extreme but very important case of materials with defects is related to highly disordered systems – amorphous states, random alloys, non-stochiometric compounds, etc. These materials have been a topic of great interest in microelectronics, e.g. non-stochiometric metal oxides are being evaluated in ReRAM device studies[13]. Many of these materials are comprised of random local arrangements of atoms and are not thermodynamically stable, requiring the use of stochastic methods to calculate their properties and to generate representative samples for simulation [14].

An area of great recent activity in MM involves low dimension systems such as free solid surfaces, thin films, material interfaces[15,16,17], motivated by continued scaling which has confounded bulk and interface effects within devices and also by the wafer level chemistry which occurs within the first few atomic layers of the surface. Addressing the problems of interest in these systems goes beyond determining static properties of materials and structures; it requires modeling dynamic processes such as the effect of deposition rate on the crystal structure of the film and how active molecules in a plasma interact with the silicon surface, etc. It also adds complexity because less can safely be assumed about the system without unphysically biasing the solution. Although it has limitations, the MD method has been indispensable tool for modeling these low dimensional systems as discussed in the following section. Many DFT packages are capable of simulating systems with the *ab initio* MD[18] algorithm; however, computational resources and the turn-around-time to complete simulations are far from what technology engineers would like to see for evaluation of multiple options or optimization of processes.

The final MM application area we will cover is simulation of nanostructures such as nanowires, nanosheets[19,20]. With the reduction of the system size, the “golden age” of being able to use MM only for the extraction of material parameters while doing the brunt of the calculation with more efficient “continuous models” which use those parameters, is waning. The MM methodology is now often applied to the entire structure and used to directly calculate macroscopic characteristics such as the current-voltage relations with methods like the non-equilibrium Green’s function (NEGF)[21] or the time dependent DFT[22]. This along with advancements to the fundamental theory which has been used for computation of ideal crystalline material properties for decades, profound progress in the development of numerical methods, software implementation, and high performance computing has advanced MM capabilities to the point of making them practical for semiconductor technology development.

3. Metals Intermixing

Metals are a key material in manufacturing high density interconnects (IC) for very large scale integration (VLSI) circuits. The design of an effective IC system is guided by many factors, among them continued scale reduction, low line resistance, minimal crosstalk, and acceptable long term reliability, e.g. mitigation of electromigration effects [23]. This multidimensional optimization results in IC systems composed of different metals and necessarily assumes contacts between them. It is a well-known effect that certain metals used in combination are susceptible to mixing caused by the process of interdiffusion[24]. This process can be intensified with pressure, applied electrical potential, and elevated temperature[25,26]. For some applications, metal interdiffusion is a desirable effect, harnessed to form a mechanically strong joint; however, for the majority of IC processing, this is not the goal. Metals with heterogenous crystal structures and foreign atoms usually increase the resistance of the contacts [17]. The mixing issue is a problem because many IC recipes require depositing thin barrier metal liner first, before the main IC metal. This liner serves as a diffusion barrier for the main contact metal, specifically to prevent its penetration into the inter-layer dielectric (ILD).[27] Mixing between the contact and liner metal would not only destroy this barrier but also increase electrons scattering from the interface, increasing overall line resistance. This is a complex system to simulate which we will discuss in subsequent sections.

3.1. Thermodynamic Considerations

Metals will intermix only if it’s energetically preferable. For many two-metal combinations, one can usually find in metallurgical textbooks or online databases, an equilibrium phase diagram[28,29] and a graph of mixing enthalpy dependence versus alloy composition to see if the metals in question form an equilibrium binary alloy and thus mix. Complications begin when the metals or their composition is not a popular entity and thus the data is absent. In this case, the mixing enthalpy needs to be calculated. There are two widely used ways to compute the equilibrium state energy of a solid. The first relies on the classical molecular dynamics (MD) method[30]. This method requires a trustworthy interatomic potential, also called a force field (FF), which may not be available for the materials of interest. Fortunately there is a universal FF that works well for metals known as EAM[31], but it requires parameters for the specific metals. If these aren’t available, a standard method for computing these consists of generating a representative set of targets for fitting, a validation suite for testing the result, and a method for optimizing, available from several optimization libraries[32]. The process of optimization itself can involve many stages, such as adding more and more targets to narrow down the set of potential parameters. The targets usually include both experimental data such as material density, elastic moduli, formation energy, etc, and data generated with a more rigorous computational method e.g. DFT[5]. Since most of the targets have error bars, the optimization can be quite complex. It’s worth noting that because of its rigor, the DFT method could be used to calculate the material properties of interest directly; however, DFT doesn’t always reproduce experimental results, even for bulk quantities such as bandgap. Because of this, using an efficient FF with its additional fitting parameters often allows more faithful matches to experiment. This fitting process is also applicable to other systems, for instances those with more than two metals or containing defects. Once the mixing enthalpy has been calculated, its sign suggests whether it’s thermodynamically preferable for two metals to mix at equilibrium conditions or exist as separate phases. The result, however, doesn’t indicate how long it would take for the materials to mix; that is where the process kinetics simulation comes into play which is the subject of the next section.

3.2. Kinetic Considerations

To evaluate the time scale of intermixing and its dependence on the initial state of the structure and process conditions, a kinetic model of the system must be developed. For this endeavor, one might be tempted to employ the same MD approach used to calculate the energy of the system as described above. However, directly integrating equations of motion for all atoms in the solid results in issue with the time scales involved. To resolve thermal vibrations of atoms, i.e. phonons, one needs to limit the time step at least by the inverse of the typical phonon frequency, which in practical simulations of solids appears to be ~10^{-2} ps. However, diffusion of atoms in solids is inherently a slow process; observable concentration changes occurs at a time scale closer to ~10^{-3} sec [33]. The result is that the MD approach becomes computationally prohibitive for modeling solid state diffusion. Another significant caveat is that the FF used in MD simulation would need to be specially fit to reproduce states with atoms far from their equilibrium positions in the crystal lattice, to capture hopping between sites, and not just for the equilibrium properties discussed in Section 3.1. To overcome these issues, the kinetic Monte Carlo (KMC) method[34] can be applied. In this method, a restricted set of physical events is selected and the appropriate rates are calculated for each of event. A Monte Carlo method is then used to sample events and advance the state of the system. In the simplest version of this method, lattice KMC (KLMC), atoms can take only fixed positions in the ideal crystal lattice. An open source implementation of the KLMC is available called SPPARKS[35]. To simulate interdiffusion, we use a customized version implemented with a model known as the binary alloy with vacancies (ABV)[36] model. In this model, a lattice site can be occupied by either atom of type A or B or remain vacant. A simple Hamiltonian, limited to only nearest neighbor interactions, is expressed as a sum of bond energies and depends on six parameters whose values can be fitted to pure metal formation energies, the energy of insertion a foreign atom or the mixing enthalpy, and the energy of vacancy formation. While rather simple, the model allows important observations about the behavior of the system. First, the possibility of mixing is directly related to the sign of AB bond energy; positive values prevent mixing. For interdiffusion to proceed, a sufficient concentration of vacancies must be assigned to the initial state, but not so that high that vacancies can coalesce and form voids. Fortunately, the final configuration is insensitive to the initial distribution of vacancies due to their high mobility. Typically the average time step is ~10^{-12}-10^{-13} sec unless the Metropolis MC algorithm[37] is selected, which effectively minimizes the system energy to reach the final configuration. These simulations don't require significant computational resources, e.g. a system of ~10^5 atoms simulating with SPPARKS using 4-8 parallel processes takes <20 min to reach the final state after ~10^8 diffusion events. Some examples of simulation results are shown in Figure 2 for the case of good mixing of Al and Cu. The pictures have been created using OVITO[38] which is a very useful tool for visualizing atomistic simulation results.

The KLMC method offers some insights into the kinetics of the intermixing although the experimental data[24] suggest that it's a much more complex process. Specifically, the main assumption in the KLMC model that atoms hop between sites in a rigid lattice is questionable. During processing, the lattice distorts and many intermediate phases of alloy are formed along the interface. Also, for technology applications, it's sometimes of interest to evaluate metals with different lattice types and imperfect crystal barrier layers, i.e. those with a grain structure. To address these problems, the off-lattice KMC model[39] has gained attention. KMC differs from KLMC in that it tracks the evolution of the system energy landscape, allowing atoms to occupy any local energy and hop through saddle-points between them. The locations of the minima and corresponding transition barriers are calculated on-the-fly using a suitable FF after every diffusion event, which makes the method extremely numerically expensive; identical systems take ~10^3 times longer to simulate with KMC versus KLMC, limiting its use [40]. The art of creating a practical KMC code is all about handling fast diffusion events[41], making full use of parallel computing[42], and avoiding barriers recalculation wherever possible[43]. As we have found at Intel, the pay-off of KMC is the quantitative agreement with experiment for interdiffusion coefficients and activation energies and the qualitative impact of lattice type, orientation, and grain boundaries. An example of its application is illustrated in Figure 3, which shows a FCC Cu - FCC Al bilayer separated with 25A of BCC Ta following

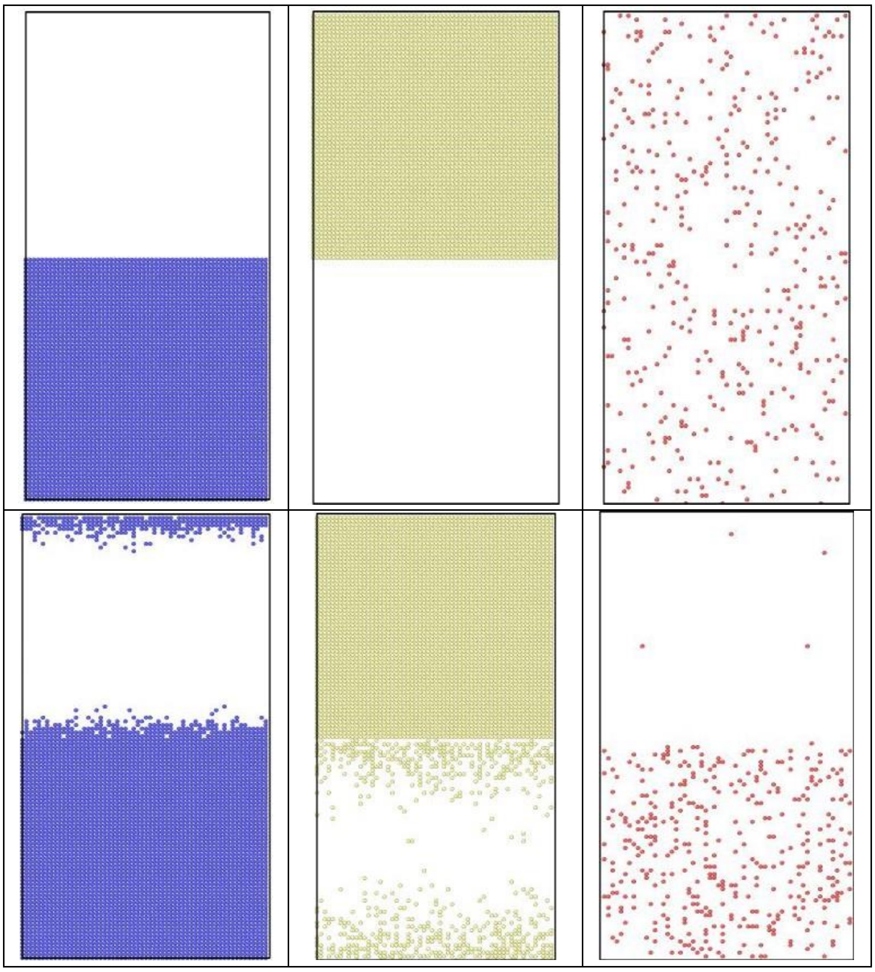


Figure 2. Initial (top 3 panels) and final (bottom 3 panels) states for Al (blue), Cu (yellow), and vacancies (red) shown left to right respectively. The final states after using KLMC to simulate 10^8 diffusion events in ~10μs. Periodic boundary conditions are set in all directions to avoid a free surface. An Cu-Al alloy forms along the interface separating initially pure metals while vacancies, which are distributed uniformly at the beginning of simulation, move into the Al region.

5 μs anneal at 700K. A bridge of Cu and Al atoms formed along the grain boundary through the barrier layer can be clearly seen after 4 days of modeling using 16 parallel processes. It's also evident that the perfect crystalline Ta is all but immiscible with both Cu and Al.

It should be noted that the MD method, despite its limitations, is being used for these type of simulations [25,26]. While certain assumptions help make these simulations more applicable, a pure metal EAM FF is not sufficient for interdiffusion simulations. The FF must include cross-type interactions of metal atoms[44] or hybridization as available in MD codes such as LAMMPS[12].

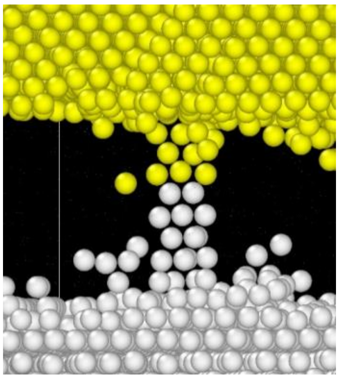


Figure 3. The bridge of Al (white) and Cu (yellow) atoms growing through Ta (hidden) separation layer along the grain boundary.

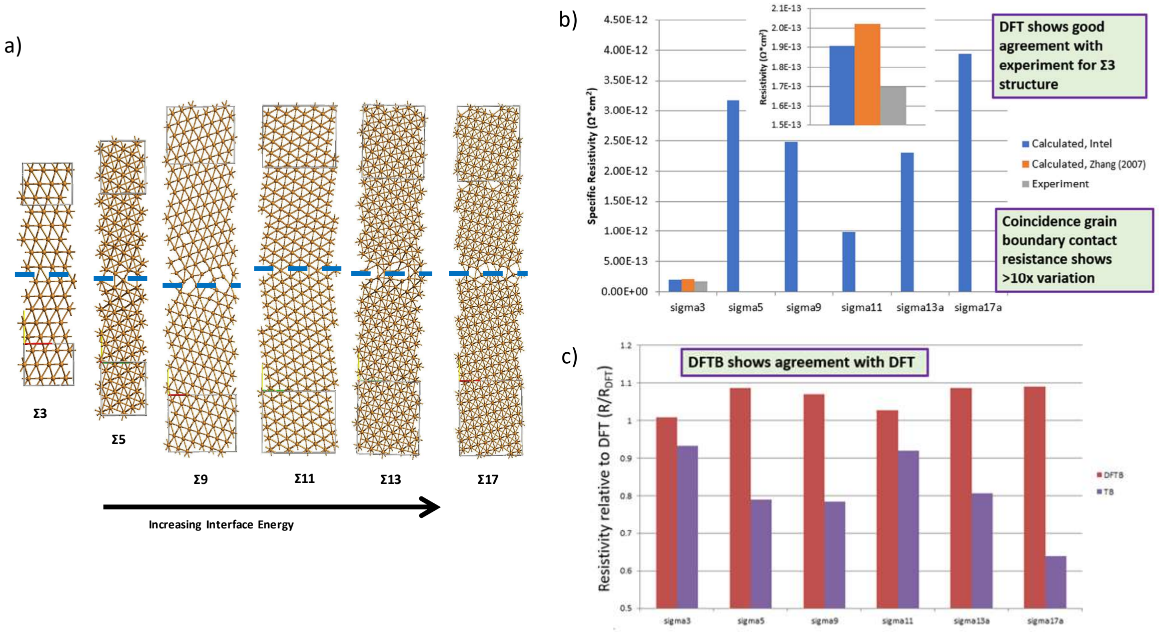


Figure 4. a) Analytically generated polycrystalline FCC Cu grain structures for different sigma grain boundaries. b) calculated resistivity for specific grain boundaries. c) comparison of TB and DFTB resistivity for different grain boundaries relative to DFT.

4. Metal Deposition for Electronic Property Calculations

The line resistance of Cu interconnects is shown to be determined by the dimensions, texture and interfaces of the metal. In this section we describe the methodology used to accurately represent the Cu microstructure in planar and trench geometries for use as metal interconnects in integrated circuits. Atomistic representations may be created analytically for bulk and planar polycrystalline configurations, followed by an energy minimization step using LAMMPS[12]. Figure 4(a) shows the generated polycrystalline representation of different grain orientations and boundary interfaces for the most stable Cu textures, generated by rotating perfect crystals to give a single grain boundary. Resistivity of the structures was calculated from using the Nonequilibrium Greens Functions framework [45]. Figure 4(b) shows the resistivities calculated using DFT simulations and compared to experimental and external reports[46]. The computationally faster method of Density Functional Tight Binding (DFTB) [47] was shown to give similar resistances as DFT simulations (see Figure 4(c)) and can be used for larger multi-grain structures with mixed grain boundary types. It is more accurate than using Tight Binding (TB) with parameters from Papaconstantopoulos [48].

Using larger analytic polycrystalline structure, over 150 configurations with roughly uniform grain sizes averaging from 2-6nm in size for 3 different lengths ranging from 7-13nm long (see Figure 5(a)) were used to calculate transmission. From these length dependent resistance plots shown in Figure 5(b), the resistivity as a function of grain size was extracted, showing smaller grains lead to higher resistivity due to increased scattering at grain boundaries. Assuming grain sizes proportional to line widths, the resistivity of interconnects including the components extracted for GB & surface scattering can be plotted as shown in Figure 5(c) showing the expected rapid increase below 5nm.

While analytic methods can be used to arrange a small amount of grains, MD can be used to simulate the deposition process, enabling the generation of truly realistic microstructures for material property calculations. Figure 6 shows the deposited microstructure results from MD simulations of Cu deposition on Ta substrates using the methodology described by Zhou and Francis[49, 50] with EAM potentials tuned for binary metal systems. Due to timestep limitations of the MD method, the deposition was simulated at an extremely exaggerated deposition rate (1 adatom/30fsec) at a temperature of 400K in order to get a sufficient thickness of 50 monolayers in the usec timeframe available. Even with the elevated conditions, the resulting (111) FCC grains with 30° rotation were in agreement with experimental reported microstructures[51].

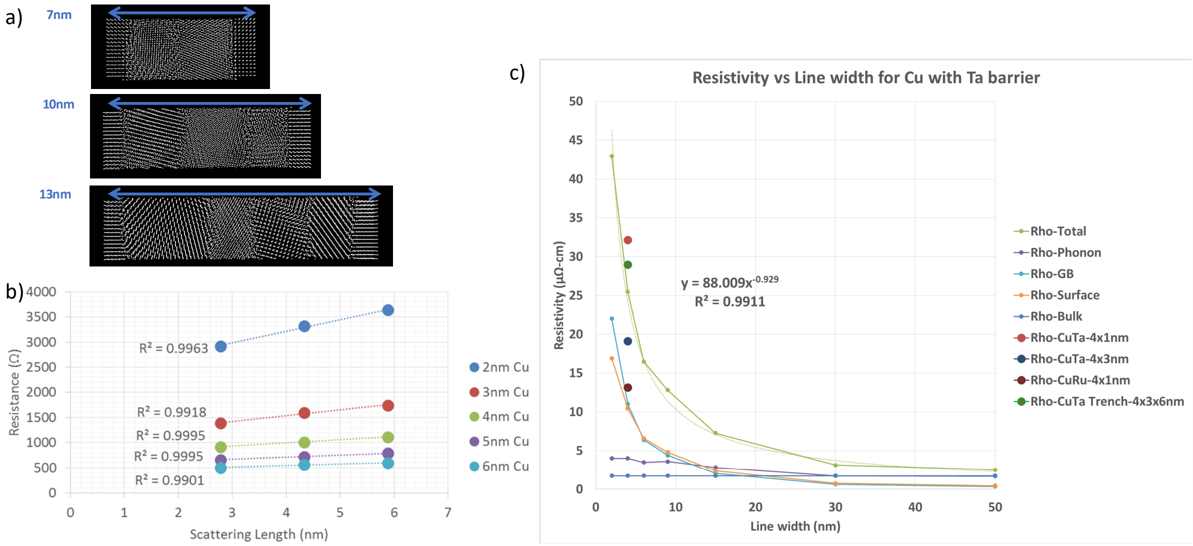


Figure 5. a) examples of analytic polycrystalline Cu structures with perfect leads used as input to DFTB transmission calculations; b) resistance curves for different lengths and average grain sizes; c) extracted resistivity curves for phonon, GB scattering and surface scattering components. Circular points are the values extracted from realistic MD deposition samples on the same plot.

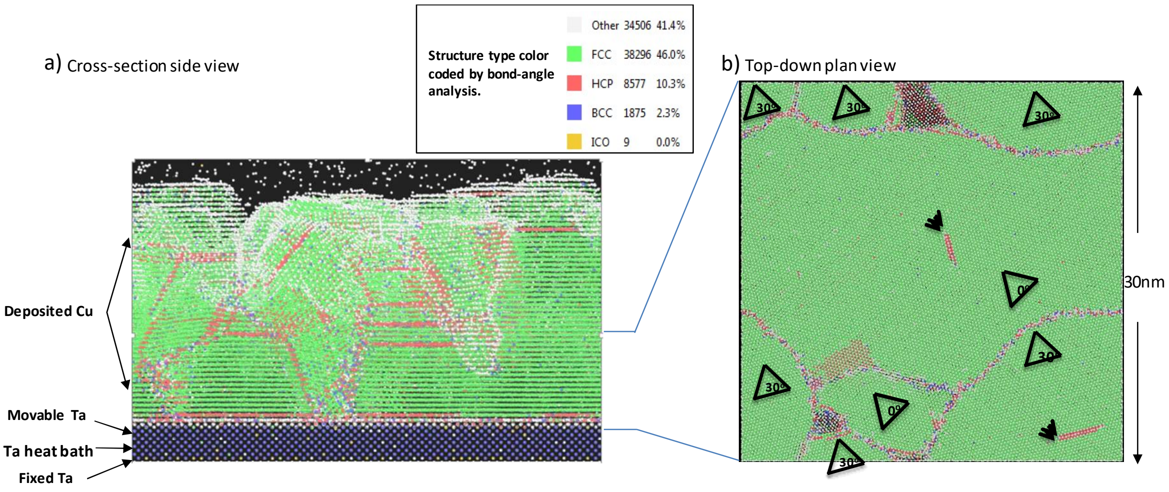


Figure 6. Cu atoms deposited on a planar Ta substrate a) side-view showing the Ta substrate layers with deposited Cu atoms; b) top-down view of a slice through the substrate and Cu deposited layer showing the microstructural grains, boundaries, and stacking faults.

Multiple instances of these microstructures were then used as input to DFTB transmission calculations to extract the resistivity for grain sizes. Extractions from slices of the realistic planar and trench MD deposition simulations are shown in the datapoints on Figure 5(c), giving good agreement with the analytic extracted curves. It confirms that the analytically generated GB structures are equivalent to the more costly full MD deposited polycrystalline ones, validating the methodology. This shows the power of using a combination of atomistic material modeling tools to generate and analyze microstructural dependence of material properties.

5. Conclusion

In this paper we briefly reviewed the state-of-the-art of MM in the context of semiconductor TCAD. We showed applications of MM approaches to problems of interest such the interaction of metals at an interface and the effect of metal grain structure on resistance. Before concluding, we would like address two vital aspects of MM. The first is

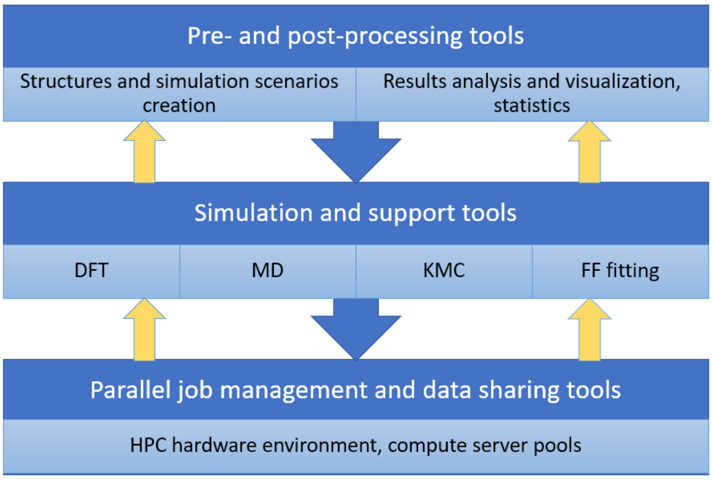

Pre- and post-processing tools
Structures and simulation scenarios creation
Results analysis and visualization, statistics
Simulation and support tools
DFT
MD
KMC
FF fitting
Parallel job management and data sharing tools
HPC hardware environment, compute server pools

Figure 7. Schematic view of material modeling software infrastructure. Downward (blue) arrows show input data flow direction, upward (yellow) arrows show simulation results flow direction.

accuracy. With enough effort, i.e., careful model selection, extensive calibration, vigilance in assuring convergence, MM can often achieve accuracy comparable to experiment. However, this level of effort is not always practical in an industrial setting, nor is it necessary. MM can still be a viable tool for assessing competing technology options provided the simulated trends are physically defensible and consistent with available data for similar systems, even when the absolute value of the results have large error bars.

The second aspect we wish to address is the framework for the ideal MM simulation environment[52]. It starts with having a tool which can create the atomistic structures of the systems we wish to model, as shown in Figure 7. This tool must be able to generate ideal as well as realistic structures, i.e. those with defects and multiple materials. Next we add reliable, highly scalable atomistic simulation code[s] to model the complete system, such as MD, KMC, or DFT and an option to seamlessly exchange atomistic structures between them. Next we would include tools for interatomic potential fitting and verification, and computational utilities for managing massively parallel jobs. To analyze the results, an extended set of postprocessing and visualization options would ideally be encapsulated into a single tool. And finally, the entire system should be connected by a flexible scripting framework, enabling construction of complex simulation flows. With such a system, a monolithic MM system could be used to simulate the majority of problems of interest versus employing individual customized flows for each application, which is the most common approach today.

In closing, we wish to recommend a recently published handbook[53] for further reading.

Acknowledgments

The authors would like to thank the members of Intel’s TCAD, J. Weber, L. Alexandrov, A, Emelyanov for their valuable contributions and discussions, suggesting ideas for MM application examples, and providing simulation results. We are grateful to Dr. Shiuh-Wuu Lee for inviting us to submit this review for publication.

References

[1] M. Bonitz, Introduction to Computational Methods in Many-Body Physics, Rinton Press Inc, (2006).
[2] E. Strohmaier, et al., “The TOP500 list and progress in High-Performance computing,” *Computer* **48**, 42–49, (2018).
[3] E. Swartz and O. Pohl, “Thermal boundary resistance,” *Rev. Mod. Phys.***61**(3), 605—668, (1989).
[4] S. Merabia and K. Termentzidis, “Thermal conductance at the interface between crystals using equilibrium and non-equilibrium molecular dynamics,” *Phys. Rev. B* 86(9), 094303, (2012).
[5] R. O. Jones, “Density functional theory: Its origins, rise to prominence, and future,” *Rev. Mod. Phys.* 87(3), 897-923, (2015).
[6] A. Abrikosov, L. Gor’kov, and I. Dzyaloshinski, Methods of Quantum Field Theory in Statistical Physics, Dover Publications, (1975).
[7] J.M. Ziman, Electrons and Phonons: The Theory of Transport Phenomena in Solids, OUP Oxford, (2001).
[8] https://en.wikipedia.org/wiki/List_of_quantum_chemistry_and_solid-state_physics_software

[9] C. Freysoldt,et al., “First-principles calculations for point defects in solids,” *Rev. Mod. Phys.* **86**(1), 253-305, (2014).

[10] B. Uberuaga and H. Jonsson, “A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths,” *J. Chem. Phys.* **113,** 9901-9904, (2000).

[11] E. Aprà et al., “NWChem: Past, present, and future,” *J. Chem. Phys*. 152, 184102 (2020).

[12] S. Plimpton, “Fast parallel algorithms for short-range molecular dynamics,” *J. Comput. Phys.* 117, 1 (1995).

[13] H. Akinaga and H. Shima, "Resistive Random Access Memory (ReRAM) Based on Metal Oxides," in *Proceedings of the IEEE* **98**(12), 2237-2251, (2010).

[14] D.A. Drabold, “Topics in the theory of amorphous materials,” *Eur. Phys. Jour. B* **68**(1), 1-21, (2009).

[15] J. Maier and H. Detz, “Atomistic modeling of interfaces in III–V semiconductor superlattices,” *Phys. Status Solidi B* **253,** 613-622, (2016).

[16] J. Schneider, et al., “ATK-ForceField: a new generation molecular dynamics software package,” *Modelling Simul. Mater. Sci. Eng.* **25**, 085007, (2017).

[17] G. Hegde, et al., “An environment-dependent semi-empirical tight binding model suitable for electron transport in bulk metals, metal alloys, metallic interfaces, and metallic nanostructures. I. Model and validation,” *J. Appl. Phys*. 115, 123703 (2014).

[18] N. Zonias, et al., “Large-scale first principles and tight-binding density functional theory calculations on hydrogen-passivated silicon nanorods,” *J. Phys.: Condens. Matter* **22**, 025303, (2010).

[19] M. Luisier, et al., “Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: From boundary conditions to strain calculations,” *Phys. Rev. B* **74**, 205323 (2006).

[20] J. Hutter, et al., “cp2k: atomistic simulations of condensed matter systems,” *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 4, 15–25, (2014).

[21] K. Stokbro K., et al., “Ab-initio Non-Equilibrium Green’s Function Formalism for Calculating Electron Transport in Molecular Devices,” in *Introducing Molecular Electronics*. *Lecture Notes in Physics,* Springer, Berlin, Heidelberg, **680**, (2006).

[22] Y. Kwok, Y. Zhang, G.-H. Chen, “Time-dependent density functional theory for quantum transport,” *Front. Phys.* **9**(6): 698–710, (2014).

[23] J. Lienig and M. Thiele, *Fundamentals of Electromigration-Aware Integrated Circuit Design*, Springer, Cham (2018).

[24] Y. Funamizu and K. Watanabe, “Interdiffusion in the Al-Cu System,” *Transactions of the Japan Institute of Metals* **12**(3), 147-152, (1971).

[25] C. Li et al., “Molecular dynamics simulation of diffusion bonding of Al–Cu interface,” *Modelling Simul. Mater. Sci. Eng.* **22**(6), 065013 (2014).

[26] M. Zaenudin, et al., "Study the Effect of Temperature on the Diffusion Bonding of Cu-Al by Using Molecular Dynamics Simulation," *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, Selangor, Malaysia, pp. 345-348, (2019).

[27] F. Zahid, et al., “Resistivity of thin Cu films coated with Ta, Ti, Ru, Al, and Pd barrier layers from first principles,” *Phys. Rev. B,* 81, 045406, (2010).

[28] T.B. Massalski, et al., “Binary Alloy Phase Diagrams,” Ed. 2, ASM International, (1990).

[29] https://www.asminternational.org/home/-/journal_content/56/10192/15469013/DATABASE

[30] M. Griebel, et al., “Numerical Simulation in Molecular Dynamics,” in *Texts in Computational Science and Engineering*, Springer-Verlag Berlin Heidelberg, 5, (2007).

[31] S. Foiles and M. Baskes, “Contributions of the embedded-atom method to materials science and engineering,” MRS Bulletin 37(5), 485-491, (2012).

[32] T. E. Oliphant, “Python for Scientific Computing,” *Computing in Science & Engineering* 9, 10-20 (2007).

[33] H. Mehrer, “Diffusion in Solids,” in *Springer Series in Solid-State Sciences*, Springer-Verlag Berlin Heidelberg, 151 (2007).

[34] A.B. Bortz, M.H. Kalos, J.L. Lebowitz, “A new algorithm for Monte Carlo simulation of Ising spin systems,” *J. Comp. Phys.* 7(1), 10-18, (1975).

[35] https://spparks.sandia.gov/index.html

[36] R. Weinkamera, yet al., “Using Kinetic Monte Carlo Simulations to Study Phase Separation in Alloys,” *Phase Transitions* 77(5-7), 433–456, (2004).

[37] N. Metropolis et al., “Equations of state calculations by fast computing machine,” *J. Chem. Phys***.** **21**(6), 1087-1091 (1953).

[38] A. Stukowski, “Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool,” *Modelling Simul. Mater. Sci. Eng.* 18, 015012, (2010).

[39] G. Henkelman and H. Jo´nsson, “Long time scale kinetic Monte Carlo simulations without lattice approximation and predefined event table,” *J. of Chem. Phys.* 15(21), 9657- 9666, (2001).

[40] L. K. Beland, et al., “Kinetic activation-relaxation technique,” *Phys. Rev. E* 84, 046704 (2011).

[41] B. Puchala, M. L. Falk, and K. Garikipati, “An energy basin finding algorithm for kinetic Monte Carlo acceleration,” *J. Chem. Phys.* 132, 134104, (2010).

[42] A. Chatterjee and D. G. Vlachos, “An overview of spatial microscopic and accelerated kinetic Monte Carlo methods,” *J. Computer-Aided Mater. Des.* 14, 253–308, (2007).

[43] J.-F. Joly, et al., “Optimization of the Kinetic Activation-Relaxation Technique, an off-lattice and self-learning kinetic Monte-Carlo method,” *J. of Phys.*: Conference Series 341, 012007, (2012).

[44] J. Cai and Y. Y. Ye, “Simple analytical embedded-atom-potential model including a long-range force for fcc metals and their alloys,” *Phys Rev B* 54, 8398-8410 (1996).

[45] G. Stefanucci and R. Van Leeuwen, “*Nonequilibrium Many-Body Theory of Quantum Systems: A Modern Introduction*,” Cambridge University Press (2013).

[46] X.G. Zhang, Kalman Varga, Sokrates T. Pantelides, “Generalized Bloch theorem for complex periodic potentials: A powerful application to quantum transport calculations” *Phys. Rev. B* **76**, 035108 (2007).

[47] P. Koskinen and V. Makinen, “*Density-Functional Tight-Binding for Beginner,s*” Computational Materials Science, **47**(1), (2009).

[48] D. A. Papaconstantopoulos, “*Handbook of the Band Structure of Elemental Solids*,” Springer (2015).

[49] X. W. Zhou and E. B. Webb III, “Atomically Engineering Cu/Ta Interfaces” Sandia Report SAND2007-5941, (2007)

[50] M. F. Francis, et al., “Atomic assembly of Cu/Ta mulitlayers: Surface roughness, grain structure, misfit dislocations, and amorphization” *Journal of Applied Physics* 104, 034310 (2008)

[51] J. S. Chawla, et al., "Electron scattering at surfaces and grain boundaries in Cu thin films and wires" *Phys Rev B* 84, 235423 (2011)

[52] D. Mejia et al., "NemoViz: a visual interactive system for atomistic simulations design," *Visualization in Engineering* **6**, 6 (2018)

[53] "Handbook of Materials Modeling," W. Andreoni and S. Yip, Ed., Springer International Publishing, (2018).

Photography & Biography

Boris A. Voinov received the M.S. degree from the Moscow Institute of Physical Engineering, in 1979, the Ph.D. degree from the State Nuclear Research Center – Institute of Experimental Physics (VNIIEF), Sarov, Russian Federation in 1989. He has been with VNIIEF from 1980 to 2003 as a Senior Research Scientist focused on theoretical, computational, and applied researches in solid state, plasma kinetics, radiation transfer, wave generation and propagation. In 2003 he joined TCAD at the Logic Technology Development, Intel Corporation.

Patrick Keys is a senior TCAD engineer at Intel Corp. He has almost 20 years of experience developing internal process modeling software and working closely with process integration teams to develop next generation transistor technologies. He holds numerous technology patents. Patrick received his B.S. degree in electronics engineering from the Univ. of Scranton, PA, a M.S. degree in Materials Science & Engr. from New Jersey Institute of Technology (NJIT), and Ph.D. in Materials Science & Engr. from the University of Florida.

Stephen Cea received the B.S. degree in electrical engineering from the University of New Hampshire, Durham, in 1990, and the M.S. and Ph.D. degrees in electrical engineering from the University of Florida, Gainesville, in 1993 and 1996, respectively. In 1996, he joined the TCAD Department, Intel Corporation, Hillsboro OR. He currently manages the Device and Process Modeling Group, TCAD Department, Intel Corporation, Hillsboro, OR. He has published over 20 works in refereed journals and conferences and holds greater than 25 patents.

Mark A. Stettler is Vice President in Technology Development and the Director of Computational and Modeling Technology at Intel Corporation. He has 5 issued patents and more than 40 publications in the field of semiconductor device modeling and process development. Mark earned a B.S. in electrical engineering from the University of Notre Dame. He also holds a master's degree and a Ph.D. in electrical engineering, both from Purdue University.

J. Microelectron. Manuf. **3**, 20030407 (2020)
doi: 10.33079/jomm.20030407

Fast and Accurate Machine Learning Inverse Lithography Using Physics Based Feature Maps and Specially Designed DCNN

Xuelong Shi, Yan Yan, Tao Zhou, Xueru Yu, Chen Li, Shoumian Chen, Yuhang Zhao*

Shanghai IC Research and Development Center, Shanghai, China 201206

Abstract: Inverse lithography technology (ILT) is intended to achieve optimal mask design to print a lithography target for a given lithography process. Full chip implementation of rigorous inverse lithography remains a challenging task because of enormous computational resource requirements and long computational time. To achieve full chip ILT solution, attempts have been made by using machine learning techniques based on deep convolution neural network (DCNN). The reported input for such DCNN is the rasterized images of the lithography target; such pure geometrical input requires DCNN to possess considerable number of layers to learn the optical properties of the mask, the nonlinear imaging process, and the rigorous ILT algorithm as well. To alleviate the difficulties, we have proposed the physics based optimal feature vector design for machine learning ILT in our early report. Although physics based feature vector followed by feed-forward neural network can provide the solution to machine learning ILT, the feature vector is long and it can consume considerable amount of memory resource in practical implementation. To improve the resource efficiency, we proposed a hybrid approach in this study by combining first few physics based feature maps with a specially designed DCNN structure to learn the rigorous ILT algorithm. Our results show that this approach can make machine learning ILT easy, fast and more accurate.

Keywords: Optimal feature maps, inverse lithography technology (ILT), deep convolution neural network (DCNN).

1. Introduction

Semiconductor industry has been progressed continuously from node to node to meet the ever increasing demand on chip performance improvement, power consumption reduction and cost reduction. The technology advancement has been enabled by various innovations in relevant fields, including new lithography exposure tools, new materials, new device architectures and new process technologies. The enormous challenges in the building of EUV lithography infrastructure has not slowed down the industry in the past, instead, the gap left by the difference in hardware resolution capability between immersion exposure tools and EUV exposure tools had created opportunities for the development and adoption of computational lithography technologies. We have witnessed the adoption of sub-resolution assist features (SRAF), multiple patterning technologies (MPT), and the source-mask co-optimization (SMO). The computational lithography technologies mentioned above have become the standard practice in developing integrated lithography patterning solutions for advanced semiconductor technology nodes. Source-mask co-optimization realizes the optimal lithography process for a selected set of patterns derived from a given set of pattern design rules. With the illumination source obtained from SMO, the lithograph process window of a chip for a design layer depends mainly on the quality of optical proximity correction (OPC) solution, which relies on the placement quality of SRAFs to a very large extent. The placement of SRAFs has gone through several evolutions, from simple rule based placement to model derived template placement, to inverse lithography technology (ILT) produced placement in hotspots fixing loop. In theory, inverse lithography has provided solid mathematical framework for achieving optimal mask solution. Although rigorous inverse lithography algorithms do exist in various forms [1, 2], full chip rigorous inverse lithography solution remains a challenging task in practice. Realization of full chip inverse lithography is not an academic interest only; it has enormous practical significance for advanced lithography process for tight pattern edge placement error control, in particular, for EUV lithography process for which stochastic effect induced edge placement error is significant. The effective way to reduce EUV

[*] Address all correspondence to Yuhang Zhao, E-mail: yhzhao@icrd.com.cn

lithography process stochastic effect is to improve image contrast through optimal assist feature placement.

The research and development in ILT has achieved fruitful progress in two directions recently. In one direction, a breakthrough has been reported in full chip rigorous mask 3D simulations through intelligent and efficient algorithm that gains computational acceleration from arrays of GPUs [3, 4]. In another direction, machine learning ILT based on deep convolution neural network (DCNN) has also been explored with success [5, 6]. Machine learning ILT is not aimed at replacing rigorous ILT entirely, instead, machine learning ILT is intended to offer sufficiently good initial ILT solution for rigorous ILT engine to take over to reach convergence with extremely fast computational speed. In essence, machine learning ILT solution can be viewed as constructing a nonlinear mapping function between the lithography target design and the rigorous ILT solution. It is not a simple point-to-point mapping; it is a function-to-point mapping. Machine learning ILT is made up of three major parts: (1) feature vector design; (2) neural network design, (3) machine learning ILT model training strategy. Feed-forward multilayer neural network architecture has been proven to possess the capability of constructing function-to-point mapping [7,8]; while convolution network has the capability of exploring spatial correlation hierarchically and extracting feature vector representation automatically through training. In semiconductor industry, DCNN has been applied to hot spot detection as a classification problem [9-12] to ILT solution as a regression problem [5, 6]. However, previous implementation of DCNN for ILT uses rasterized lithography target design as input, with such pure geometrical image as input, the feature vectors extracted from DCNN lack of intuitive physical interpretation, they cannot address the critical questions regarding feature vector design, i.e., the feature vector resolution, the feature vector sufficiency, and the feature vector efficiency. The optimality of the feature vector extracted from such DCNN implementation is much more sensitive to the training samples selected.

In our previous reports, we have presented our machine learning OPC and machine learning ILT results based on physically derived feature vector design followed by a shallow (5 to 6 layers) feed-forward neural network [13, 14]. For machine learning ILT with our proposed physically derived feature vector design, the feature vector length needs to be around 140 to achieve satisfactory model accuracy, which will demand considerable memory resource in practical implementation. To lift the memory resource burden while still taking advantage of physics based feature vector design, we propose a hybrid approach in this study, which uses first few physics based feature maps as input, followed by a specially designed DCNN. The specially designed DCNN possesses the desired properties of being wide receptive field and of being able to preserve high resolution. It turns out that this hybrid approach can make machine learning ILT easy, fast and more accurate.

2. Feature Vector Design for Machine Learning ILT

Machine learning based ILT can be generally stated as: *For a given ADI target layer and a fixed optimal mask generation mechanism (illumination source + mask type + rigorous ILT algorithm), there should exist a unique mapping function between ADI target data and ILT data, as shown in Figure 1.*

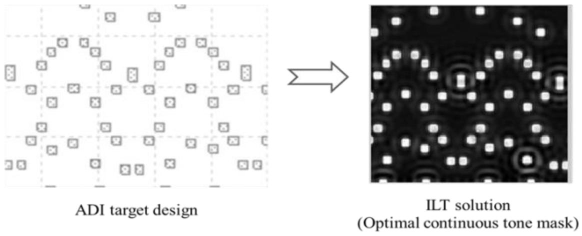


Figure 1. Mapping from ADI target to ILT image.

Mathematically, it can be expressed as:

$$\text{ILT } function\,(x, y) = F\,(ADI\ target\ patterns\,(x, y)) \quad (1)$$

As we emphasized earlier, it is not a point-to-point mapping, it is a function-to-point mapping. *The value of ILT solution at point (x, y) not only depends on the value of ADI target data at point (x, y), but also depends on all values of ADI target data around the point (x, y) within an influence range.* Before we proceed to address the question of how to design feature vector to describe the neighboring environment around a point (x, y), we should first ask the question: *how many* degrees of freedom does the neighboring environment around a point (x, y) have? The theoretical answer is: the degree of freedom of the neighboring environment around a point (x, y) is infinite. Therefore, a complete description of the neighboring environment around a point (x, y) is impossible. Fortunately, a description with infinite resolution is often not required practically. This is true for machine learning based computational lithography, because the imaging system used in lithography process does not possess infinite resolution. This fact suggests that the number of effective degree of freedom of the neighboring environment around any point (x, y) can be considered finite practically. *This observation and fact is the very foundation of all computational lithography*. The second question we need to address is: what is a feature vector and what desired properties a feature vector should have? Essentially, a feature vector is a mathematical representation that describes the neighboring environment around a point (x, y) in a quantitative way. *As a measurement device, a feature vector must address the following important properties, i.e., the measurement resolution, the measurement sufficiency (completeness), and the measurement efficiency.* In addition, it is very desirable for a feature vector to possess a property such that the mapping function from input to output of the neural network model is less nonlinear and smooth (differentiable), or even monotonic (hopefully).

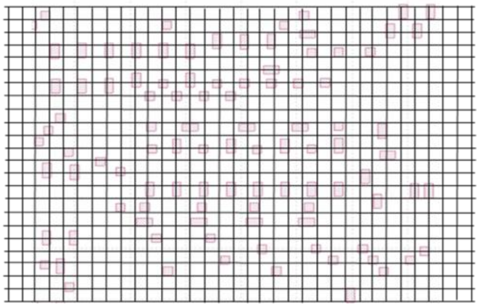


Figure 2. Divide the neighboring environment into cells.

To elucidate the concept of *measurement resolution and measurement efficiency of a feature vector*, we can look at Figure 2. To describe the neighboring environment around a point (x, y), we can divide the influencing area into small cells. Assume the influencing range is 1.0 μm each side, and the cell size is *x* nm, then the cell size *x* determines the resolution of the feature vector representation, and the total number of cells = $(2\cdot1000/x)^2$ represents the maximum length of the feature vector for a complete description with resolution *x* nm. Clearly, the smaller the cell size x, the higher the measurement resolution; and the higher the resolution of the feature vector representation, the longer the feature vector is. To serve the machine learning based ILT properly, the resolution of the feature vector representation must meet a minimum requirement, which is determined by lithography process imaging condition, i.e., cell size $x = k\cdot\lambda/(NA(1+\sigma_{max}))$. The k coefficient is related to the degree of spatial coherence of the illumination, which depends on the effective illumination area of the source. A typical cell size for high NA immersion lithography process is around 15nm to 20nm, therefore, the estimated feature vector length for a complete description is $(2000/20)^2 = 10000$. Of course, such a simple and plain encoding scheme for neighboring environment lacks of efficiency, because the encoding scheme does not explore the characteristics of the lithography process, it treats all cells equally and independently, it does not explore all symmetry properties among all the cells. Intuitively, not every cell has the same influence on the point of interest, on average, the closer the cell to the point of interest, the more important the cell is. As to the *sufficiency of a feature vector*, it is related to the capability of the feature vector in describing the neighboring environment completely within allowed error bound. Simply stated, for any two feature vectors $\mathbf{X}_1$, $\mathbf{X}_2$, if $\mathbf{X}_1= \mathbf{X}_2$, then, the condition $|F(\mathbf{X}_1) - F(\mathbf{X}_2)| \le \varepsilon$ (ε is the allowed error bound related to data noise) CANNOT be violated.

There have been several reported ways of designing feature vectors for computational lithography. Incremental concentric square sampling [15], incremental concentric circle area sampling [16], polar Fourier transform [17] have all been proposed to be used for constructing feature vectors for computational lithography. These feature vector designs do not address the optimality of the designed feature vector, and most of them are pure geometrical based feature vectors, except the design based on polar Fourier transform. Feature vectors based on "*geometrical rulers*" have intrinsic deficiency in machine learning computational lithography; this is particularly true for inverse lithography which grows assist features out of blank areas in mask space. As it is known, rule based assist feature insertion based on geometrical measurement has abrupt change points in the rule table. Therefore, machine learning inverse lithography using "geometrical ruler" based feature vector as neural network input must possess more complicated network structure to learn those abrupt change points in order to map the feature vector into correct response function domain. Feature vectors derived from polar Fourier transform made progress by exploring the characteristics of the lithography process partially, however, it still fails to fully take the imaging process physics into account. Feature vector design is essentially an information encoding scheme design. For machine learning computational lithography, there are three spaces we can use for information encoding, the lithography target space, which is pure geometrical; the mask space, which has geometrical information and optical property information; the image space, which contains information about design geometries, mask optical properties and imaging formation characteristics. From an information point of view, information in lithography target space is not complete (without specifying optical properties of the background and the pattern covered areas), if feature vector design is in lithography target space, then the subsequent DCNN must learn mask optical properties, nonlinear imaging formation process and rigorous ILT algorithm. Information in mask space is complete and of highest resolution. If feature vector design is in mask space, then the subsequent DCNN must learn nonlinear imaging formation process and rigorous ILT algorithm. Information in imaging space can be used to recover information in mask space fully within the resolution limit defined by optical imaging condition. If feature vector design is in image space, then the subsequent DCNN *only need* to learn the rigorous ILT algorithm. Between mask space and image space, which space is narrower in terms of encoding efficiency? In mask space, the "function space" is constrained by design rules of the layer; while in image space, the "function space" is constrained by both design rules and imaging conditions. Stated explicitly, *all aerial images derived from a given imaging condition constitute a special class of functions.* In other words, the "function space" in image space is narrower than the "function space" in mask space, and the information lost in image space in comparison with that in mask space is beyond the optical imaging resolution. *Therefore, optimal feature vector design for computational lithography should be related to optimal and efficient representation of aerial images of the class at hand.*

Now the question becomes how to represent aerial images most efficiently? The aerial image function I(x,y) is a band-limited function. While a real function with finite bandwidth Ω can always be represented by a set of basis functions of the same bandwidth, there still exists the question whether there is an optimum set of basis functions among all the possible sets of basis functions with bandwidth, Ω. By the optimum set of basis functions, it means that only the minimum number of the basis functions that are needed to approximate any real valued function of bandwidth, Ω, for a specified error requirement. To seek the optimal representation of aerial image function, we can refer to the imaging equation of Hopkin's, which can be decomposed into a sum of coherent imaging system for partially coherent illumination, as shown in Equation (2) below.

$$I(x,y)=\sum_{i=1}^{\infty}\alpha_i\left|\varphi_i\otimes M\right|^2 \qquad (2)$$

Where ⊗ represents the convolution operation between the ith kernel and the mask transmission function M. {ϕ_i} and {α_i} are the set of eigenfunctions and eigenvalues of the transmission cross coefficients matrix (TCCs). This optimal imaging system decomposition is originally designed for fast aerial image calculation under partial coherent illumination, and it has been proved that this decomposition scheme is the optimal decomposition in terms of computational efficiency [18]. From an information theory point of view, we can interpret it as an optimal and most efficient aerial image information encoding scheme. This suggests that imaging system kernels {ϕ_i} captures imaging system characteristics fully, and they are a set of natural and optimal "*optical rulers*" for measuring or estimating the neighboring environment around a point (x, y), because the set of {ϕ_i} eigenfunctions are orthonormal functions. Based on the above reasoning, we define {S_1, S_2, … , S_N} as the feature vector, with S_i being defined as:

$$\mathrm{S_i}=\left|\phi_i\otimes M\right|^2 \qquad (3)$$

Then, the machine learning inverse lithography problem can be reformulated from Equation (1) to Equation (4).

$$\text{ILT } \mathit{function}(x, y) = \mathcal{F}(S_1\,(x, y), S_2\,(x, y),\ldots S_N\,(x, y)) \qquad (4)$$

The idea of using imaging eigen signal set {S_i} to describe aerial image has been used previously for OPC model and lithography two-dimensional patterns' quantification [19, 20]. Now we turn to the question of how to obtain the approximate function *F*, this is related to neural network design.

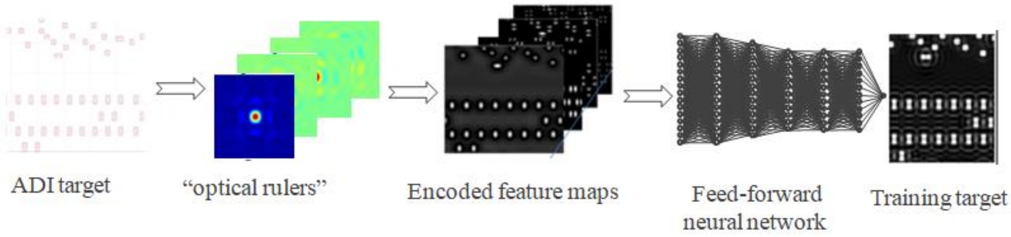


Figure 3. Feed-forward neural network model.

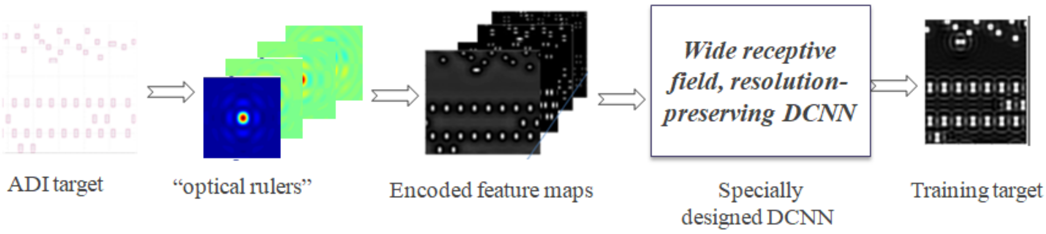


Figure 4. Hybrid approach machine learning inverse lithography model.

3. Machine Learning Based ILT and Results

With feature vectors calculated using Equation (3), a general mapping function described by Equation (4) can be constructed using a feed-forward neural network structure, as suggested by the universal approximation theorem [7, 8]. The results based on this approach have been reported in our previous report [14]. Figure 3 shows the key elements of the approach.

Since both the input feature vector maps and the output (continuous tone mask) are band-limited functions, they are smooth and differential functions. This property makes the mapping function construction easier using feed-forward neural networks. However, we found that the required feature vector is still considerably long in size (140 elements in our study) in order to achieve good model. This will impose considerable requirement on memory resource in practical applications. To ease the memory resource requirement while keeping physics based feature vector as input, we have taken a hybrid approach in this study. In this hybrid approach, we used $\{S_1, S_2, S_3, S_4, S_5\}$ five feature maps as input into a specially designed deep convolution neural network (DCNN). The basic idea is to use first few physics based feature maps, which are supposed to be able to provide sufficient information to represent mask optical properties and imaging process characteristics, then the subsequent DCNN to develop more deeper and efficient representation for ILT modeling and to accomplish coordinated regression. This is because both input feature maps and the output image (continuous tone mask) have certain degree of spatial correlation, i.e., neighboring pixels are correlated. To serve machine learning inverse lithography purpose, the specially designed DCNN structure should possess certain desired properties: (1). The wider the receptive field, the better, in order to explore the spatial information around a point (x, y); (2). The original resolution of the image should be preserved; (3). The depth of the DCNN should be moderate, so that there will be no need to have residual connections in the network structure for easy training. Following these design guidelines, we replace all pooling layers with bath normalization layers, and we use ReLU as the activation function. The convolution kernels are all 3x3 in size, and the stride step size is 1. The design of our hybrid approach is shown in Figure 4.

The training of the neural network model needs to include training samples and test samples, and they are selected from the periphery areas of a 28nm SRAM design via layer. The pattern selection strategy is the same as that for OPC model calibration and SMO. Total number of images for training is 134, and total number of images for model test is 48. We have tried both He initialization and orthogonal initialization for weights in model training, and we found there is no essential difference in terms of the model quality

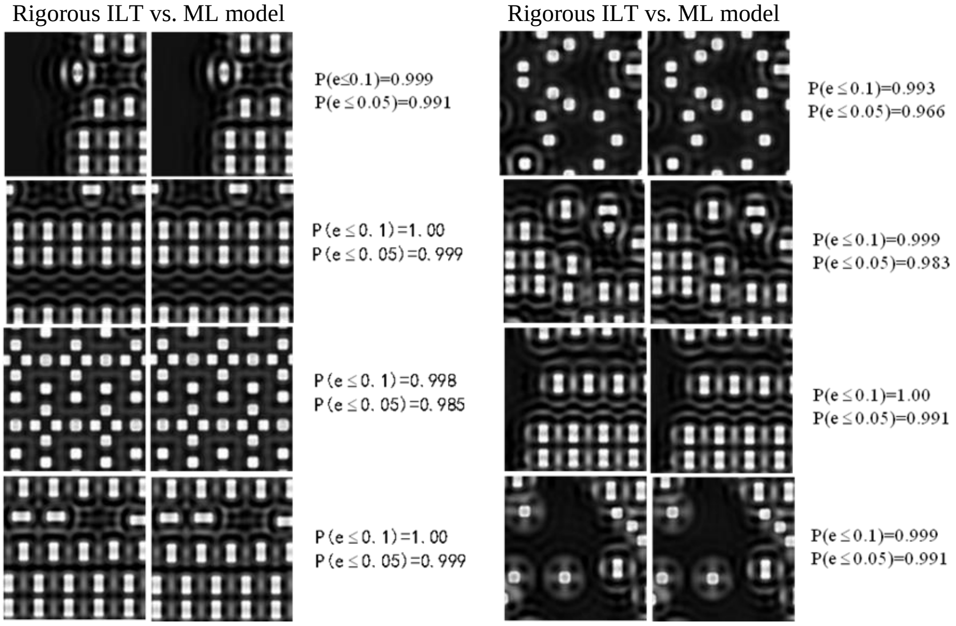


Figure 5. Images from rigorous ILT solutions and from machine learning model for training set, model input: {S1:S5}.

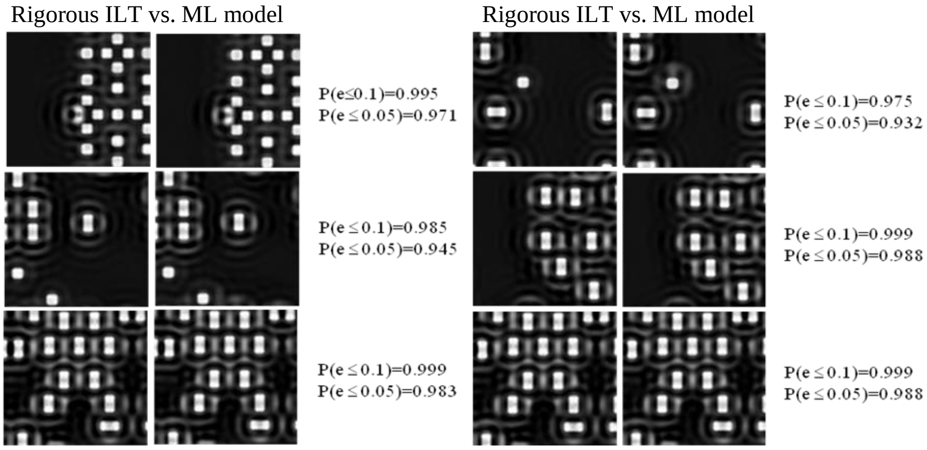


Figure 6. Images from rigorous ILT solutions and from machine learning model for test set, model input: {S1:S5}.

from these two different weight initialization schemes. The learning rate used is 5×10^{-5}, and Adam optimizer is used in training.

To assess the model quality, we first normalize the rigorous inverse lithography solution into [0, 1] using a common normalization factor, then we use two metrics to quantify the quality of a model. Let O denote the normalized rigorous inverse lithography solution image, and $\hat{O}$ the neural network model predicted image. Then the first metric we used is the probability $P(|O - \hat{O}| \leq \varepsilon)$ where $\varepsilon = 0.1$ and 0.05, and the other metric used is RMSE. For comparison purpose, besides using $\{S_1, S_2, S_3, S_4, S_5\}$ as DCNN input, we also used {Aerial image} and {Aerial image + S_1:S_5} as DCNN input. The model training error statistics and test error statistics are shown in Table 1 below.

The visual comparison between images from rigorous ILT solutions and from our machine learning model for training set and test set are shown in Figure 5 and Figure 6.

As it can be seen from Table 1, the first five feature vector maps (images) $\{S_1$:$S_5\}$ are better model input design than aerial image alone. Aerial image is the weighted sum of many signals (images) from independent imaging formation kernels $\{\phi_i\}$, as expressed in Equation (2). The sum operation makes the original information collapse to a certain extent, the set of independent feature vector maps (images) $\{S_1$:$S_5\}$ preserves the original information better.

Table 1. Model training error statistics and verification error statistics

Model input	Error spec.	$P(\|O -\hat{O}\| \le \varepsilon)$		RMSE (x10^{-4})	
		Training set	Test set	Training set	Test set
Aerial images	$P(\|O -\hat{O}\| \le 0.10)$	0.987	0.976	3.5	4.4
	$P(\|O -\hat{O}\| \le 0.05)$	0.928	0.916		
$\{S_1:S_5\}$	$P(\|O -\hat{O}\| \le 0.10)$	0.999	0.995	1.8	2.6
	$P(\|O -\hat{O}\| \le 0.05)$	0.989	0.968		
Aerial images + $\{S_1:S_5\}$	$P(\|O -\hat{O}\| \le 0.10)$	0.998	0.989	1.9	2.9
	$P(\|O -\hat{O}\| \le 0.05)$	0.987	0.965		

With the first five feature vector maps (images) $\{S_1:S_5\}$ as DCNN input, $P(|O - \hat{O}| \le 0.05)$ can reach 96.8%. This is better than the model performance using feed-forward neural network with long feature vector (feature vector length =140), the feed-forward neural network model can only achieve $P(|O - \hat{O}| \le 0.1) = 99.0\%$ and $P(|O - \hat{O}| \le 0.05) = 87.5\%$. The improved model accuracy of the hybrid approach proposed in this study may result from a combination of the physics based feature maps, which contain information about the image formation mechanism, and the power of DCNN, which possesses the great capability of further exploring spatial information from $\{S_1:S_5\}$ and of constructing deeper representation most suitable for learning rigorous ILT mechanism.

Besides the greatly improved model accuracy in comparison with the feed-forward model, the speed enhancement relative to rigorous ILT is also significant. With 4 CPUs (Intel Xeon E7-8855-V4, 2.1 GHz, each CPU has 14 cores), it takes 12.1 seconds on average for a 20μmx20μm patch. In comparison with rigorous algorithm (assume 100 iterations for reaching convergence), the estimated speed gain factor is about 25 or more. By running the model on a single GPU (Nvidia telsa M60), additional speed enhancement by a factor of 20 can be achieved.

4. Conclusion

Inverse lithography technologies can theoretically provide the ultimate optimal mask solutions once the lithography process imaging condition is fixed. However, its full chip implementation has been in stagnation for a long time due to its lack of sufficient speed using rigorous algorithms. A hybrid approach by combining machine learning inverse lithography technology with faster rigorous ILT algorithms has paved the way for its full chip implementation. Due to high accuracy requirement, machine learning inverse lithography is not intended to provide the final ILT solution entirely; rather, it provides a sufficiently good initial solution for a rigorous engine to take over and to achieve final converged solution with very few iterations. In our proposed machine learning inverse lithography method, we use information in image space directly instead of information in design geometrical space as model input to lift the burden for the model to learn very non-linear imaging physical process. We also employ a specially designed DCNN that can both develop more efficient representation for machine learning ILT from imaging space information and do coordinated regression. The new innovative method has made machine learning ILT easy, fast and more accurate.

References

[1] A. E. Rosenbluth, S. Bukofsky, C. Fonseca, M. Hibbs, K. Lai, R. N. Singh, and A. K. Wong, “Optimal mask and source patterns to print a given shape”, J. Microlith. Microfab. Microsys. I(1), 13-30 (2002).

[2] L. Pang, Y. Liu, and D. Abrams, “Inverse lithography technology (ILT): a natural solution for model-based SRAF at 45nm and 32nm”, Proc. SPIE, 6607, (2007)

[3] Yeung, M. and Barouch, E., “Development of fast rigorous simulator for large-area EUV lithography simulation.” Proc. SPIE 10957,109571D, (2019).

[4] L. Pang, E. V. Russell, B. Baggenstoss, M. Lee, etc., “Study of mask and wafer co-design that utilizes a new extreme SIMD approach to computing in memory manufacturing – full chip curvilinear ILT in a day”, Proc. SPIE 11148, (2019).

[5] Shibing Wang et al., “Efficient full-chip SRAF placement using machine learning for best accuracy and improved consistency”, Proc. SPIE 10587, (2018).

[6] Song Lan, Jun Liu, Yumin Wang, Ke Zhao, Jiangwei Li, “Deep learning assisted fast mask optimization.”, Proc. SPIE 10587, (2018).

[7] Cybenko, G., “Approximation by superposition of a sigmoidal function.”, *Mathematics of Control. Signals and Systems.* 2, (1989), 303-314.

[8] Hornik, Kurt, “Approximation Capabilities of Multilayer Feedforward Networks.”, *Neural Networks*, Vol. 4, (1991), 251-257.

[9] Haoyu Yang, Luyang Luo, Jing Su, Chenxi Lin, and Bei Yu, "Imbalance Aware Lithography Hotspot Detection: A Deep Learning Approach.", Proc. SPIE 10148, (2017).
[10] Yibo Lin, Xiaoqing Xu, Jiaojiao Ou, David Pan, "Machine Learning for Mask/Wafer Hotspot Detection and Mask Synthesis", Proc. SPIE 10451, (2017).
[11] Tetsuaki Matsunawa, Shigeki Nojima, and Toshiya Kotani, "Automatic Layout Feature Extraction for Lithography Hotspot Detection Based on Deep Neural Network.", Proc. SPEI 9781, (2016).
[12] Yiwei Yang, Zheng Shi, Litian Sun, Ye Chen, Zhijuan Hu, "A kernel-Based DfM Model for Process from Layout to Wafer", Proc. SPIE 7641, (2010).
[13] Xuelong Shi, Yuhang Zhao, Shoumian Chen, Ming Li, "Optimal feature vector design for computational lithography", Proc. SPIE 10961, (2019).
[14] Xuelong Shi, Yuhang Zhao, Shoumian Chen, Chen Li, "Physics based feature vector design: a critical step towards machine learning based inverse lithography", Proc. SPIE 11327, (2020).
[15] Gu A. and Zakhor A., "Optical proximity correction with linear regression.", *IEEE Trans. Semicond. Manuf.* Vol. 21, (2008), 263-71.
[16] Tetsuaki Matsunawa, Bei Yu and David Pan, "Optical proximity correction with hierarchical Bayes model.", *Proc. SPIE* 9426, (2015).
[17] Suhyeong Choi, Seongbo Shim, Youngsoo Shin, "Machine learning (ML)-guided OPC using basis functions of polar Fourier transform.", Proc. SPIE 9780, (2017).
[18] Y. C. Pati nd T. Kailath, "Phase-shifting masks for microlithography: automated design and mask requirements", J. Opt. Soc. Am. A. vol. 11 (9), 1994.
[19] Xuelong Shi, Tom Laidig, J. Fung Chen, Doug Van Den Broeke, Stephen Hsu, Michael Hsu, Kurt Wampler, Uwe Hollerbach, "Eigen decomposition based models for model OPC", Proc. SPIE 5446, (2004).
[20] Xuelong Shi, J. Fung Chen, Doug Van Den Broeke, Stephen Hsu, Michael Hsu, "Quantification of two-dimensional structures generalized for OPC model verification", Proc. SPIE 6518, (2007).

Photography & Biography

Xuelong Shi is a technologist in Shanghai IC R&D Center. He has a BS degree in chemistry from University of Science and Technology of China in 1985, and a PhD degree in physical chemistry from Columbia University in 1993. He started his career in the field of lithography in 1996, and has since been working on lithography process, illumination optimization, source-mask-co-optimization, optical proximity correction (OPC). His current interests are machine learning based computational lithography and machine learning based computational metrology.

Yan Yan graduated from Department of Electronic Engineering, Xiamen University with a BS degree in 2015 and obtained her MS degree from Department of Micro-electronics, Shanghai Jiao Tong University in 2019. She joined Shanghai IC R&D Center (ICRD) in 2019 and is currently working on machine learning computational lithography and machine learning metrology.

Tao Zhou graduated from School of Microelectronics and Solid State Electronics, University of Electronic Science and Technology of China in 2009 and obtained his PhD at Shanghai Institute of Microsystems and Information Technology, Chinese Academy of Sciences in 2014. He joined Shanghai IC R&D Center (ICRD) in 2018, he engaged in CMOS image sensor (CIS) imaging optimization and image analysis. At present, he mainly studies the measurement and characterization of SEM images from lithography and focuses on various image processing problems in fab manufacturing. He has published more than 30 papers in international journals such as Advance Sciences, Scientific Reports, IEEE Photonics Technology Letters, etc, and have more than 30 patents in relevant fields.

Xueru Yu received the B.S. degree in Microelectronics from Shanghai Jiao Tong University, Shanghai, China in 2015 and the Master degree in Integrated Circuit Engineering from Shanghai Jiao Tong University, Shanghai, China in 2018. He joined Shanghai IC R&D Center (ICRD) in 2018. He works in data mining and computer vision based on artificial intelligence.

Chen Li received the B.S. degree in physics from Peking University, Beijing, China, in 2001. He received the Ph.D. degree at the Institute of Microelectronics, Peking University. He was a visiting scholar at the Department of Electrical Engineering, Columbia University, U.S.A. in 2008. Now he is with Shanghai Integrated Circuits R&D Center Ltd. (ICRD), Shanghai, China. He is the director of AI technology department in ICRD. His research interests include AI chip, AI algorithms and their applications in IC industry.

J. Microelectron. Manuf. **3**, 20030408 (2020)
doi: 10.33079/jomm.20030408

Machine Learning based Optical Proximity Correction Techniques

Pengpeng Yuan [1], Taian Fan [1], Yaobin Feng [3], Peng Xu [1, *], Yayi Wei [1, 2, **]

[1]*Institute of Microelectronics, Chinese Academy of Science, Beijing, China, 100029*
[2]*University of Chinese Academy of Science, Beijing, China, 100029*
[3]*Yangtze Memory Technologies Co., Ltd, Wuhan, Hubei, China, 430000*

Abstract: The shrinking of the size of the advanced technological nodes brings up new challenges to the semiconductor manufacturing community. The optical proximity correction (OPC) is invented to reduce the errors of the lithographic process. The conventional OPC techniques rely on the empirical models and optimization methods of iterative type. Both the accuracy and computing speed of the existing OPC techniques need to be improved to fulfill the stringent requirement of the research and design for latest technological nodes. The emergence of machine learning technologies inspires novel OPC algorithms. More accurate forward simulation of the lithographic process and single turn optimization methods are enabled by the machine learning based OPC techniques. We discuss the latest progress made by the OPC community in the process simulation and optimization based on machine learning techniques.

Keywords: Optical Proximity Correction, Machine Learning, Deep Learning, Lithography.

1. Introduction

Optical proximity correction (OPC) becomes critical for the process of current advanced technological nodes. The conventional methods of the optical proximity correction rely on the empirical rules or the combination of the parametric models and traditional optimization methods most likely in the iterative sense. The empirical rules highly depend on the experience of process engineers and works well for the early technological nodes with larger critical dimensions, but the model based methods are required for the sub 100nm technological nodes. Considering the difficulty of the rigorous mathematical simulation of the physical and chemical process involved in the optical lithographic process, simplified models with empirical parameters are usually applied in the actual OPC process[1, 2]. Even the contemporary numerical methods such as Finite Difference Time Domain (FDTD)[3–5] or Finite Element Method (FEM) *et. al.* are able to provide more accurate solutions to the optical imaging process, photo-chemical reactions and so on, the formidable cost of computation power hinders the application of such methods to larger scale systems such as full chip level optimization problems. The past efforts in deriving more accurate analytical or semi-numerical models for the forward simulations of relevant physical chemical processes[6,7] boost the development of more practical and efficient OPC technologies. Even the complexity of the photo-resist reactions slows down the progress of obtainment of more reliable resist model, and so does the etch model, the optical imaging problems can usually be well resolved after simplified assumptions are made to the optical imaging systems. The appearance of the computational tractable models makes the iterative optimization of the mask shapes possible which becomes the corner stone of the current OPC technologies. Combining with the latest setup of the optical lithography machines which enables the variable illumination conditions[8, 9], source-mask optimization (SMO)[10–12] also becomes an important part of the OPC workflow. Until the techniques of the insertion of sub-resolution assist features (SRAF) being added to the arsenal of OPC toolkit to enlarge the process window of the optimized mask patterns, the framework of contemporary OPC is settled down. However, the room of the improvement of current OPC workflow remains and the rise of data science as well as machine learning provides huge amount of opportunities for the computational lithography community.

2. Optical Lithographic System

The optical microlithography system mainly includes four parts: source, mask/reticle, exposure system and wafer. To avoid the inhomogeneity of

[*] Address all correspondence to Peng Xu, E-mail: xupeng2019@ime.ac.cn
[**] Address all correspondence to Yayi Wei, E-mail: weiyayi@ime.ac.cn

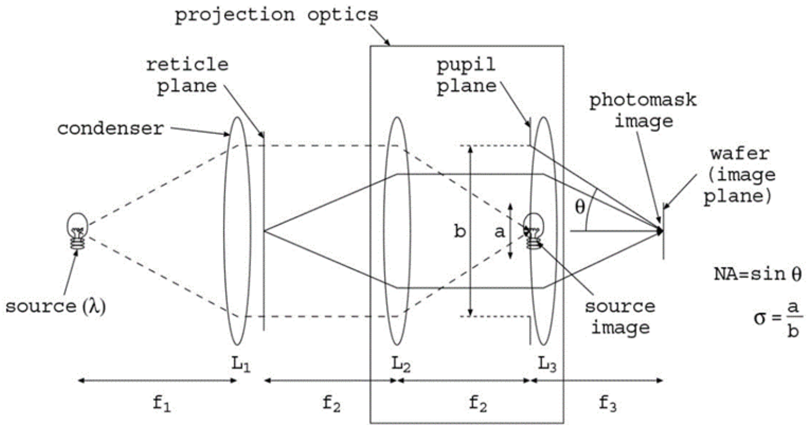


Figure 1. Illustration of an optical lithographic projection system[13].

illumination on the photomask, Kohler's method of illumination is applied. The source or the image of the source is placed at the focal plane of the condenser. The photomask/reticle is then illuminated by the parallel beam and the energy distribution on the top plane of the mask is then homogenous in the idealistic situation. Traditional mask is the binary intensity mask. They are formed by the chromium on glass. Different types of fused silicon are applied for varied illumination wavelengths. Phase shift masks are also introduced to improve the image quality. The image of the scattering sources on the photomask is formed on the wafer after the projection optics. The standing wave pattern is formed by the reflection from the photoresist/wafer interface. There are two types of photoresist: positive resist and negative resist. They response differently to the illumination.

3. Machine Learning Based OPC

The machine learning and data driven perspective may change the OPC workflow mainly in three aspects. Firstly, more accurate and fast models become accessible after the introduction of novel tools such as deep neural networks as a good universal approximator which should be beneficial even they are simply embedded into the traditional OPC framework; Secondly, the expensive and time consuming iterative optimization process of prevailing OPC techniques may be replaced by the single run computation of well trained models which directly perform the optimization process including the mask optimization (MO), SMO, SRAF insertion and so on; Third, the whole workflow of the OPC may be modified by the data driven methodology and the changes may not be constrained within the scopes of feature pattern selection and hotspots detection. The novel full chip level solution may be enabled in the future. We shall discuss the recent progress of the machine learning based OPC technologies in the three directions mentioned above separately.

3.1. Negative Tone Development

The simulation of the lithography mainly contains three parts: Optical model, Resist model and Etch model. The first two process is coarsely shown in Figure 1.

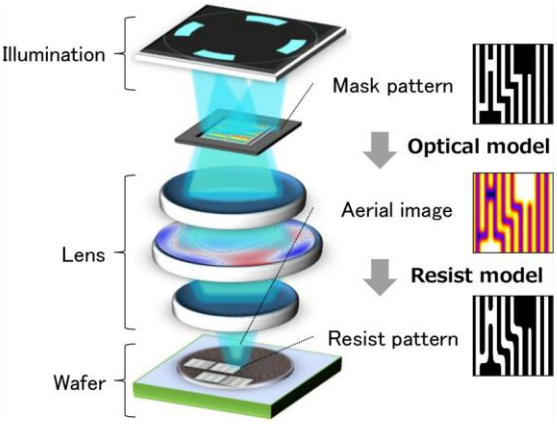


Figure 2. Lithography simulation[14].

The ideal situation of the forward simulation is the success of the *ab initio* calculation. For the optical imaging system, this target is more achievable. The realistic imaging system is usually simplified and mathematical abstraction can be done within the framework of optics[15]. In Figure 2, a typical optical configuration is shown. The aerial

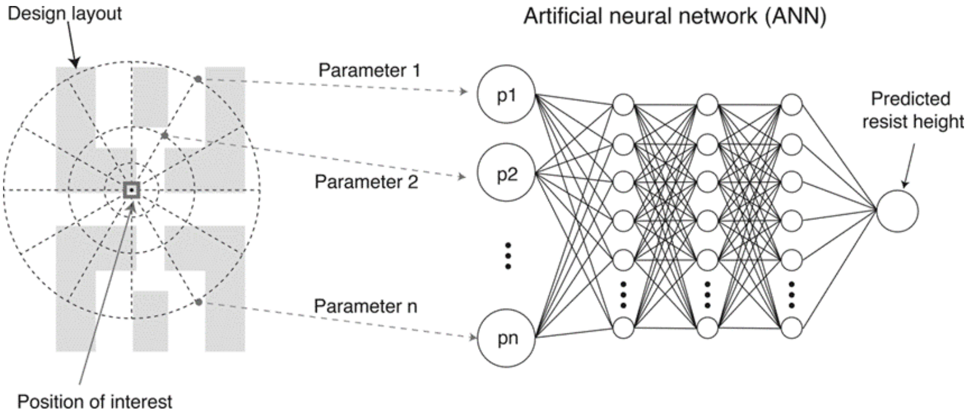


Figure 3. ANN-based 3D resist model[21].

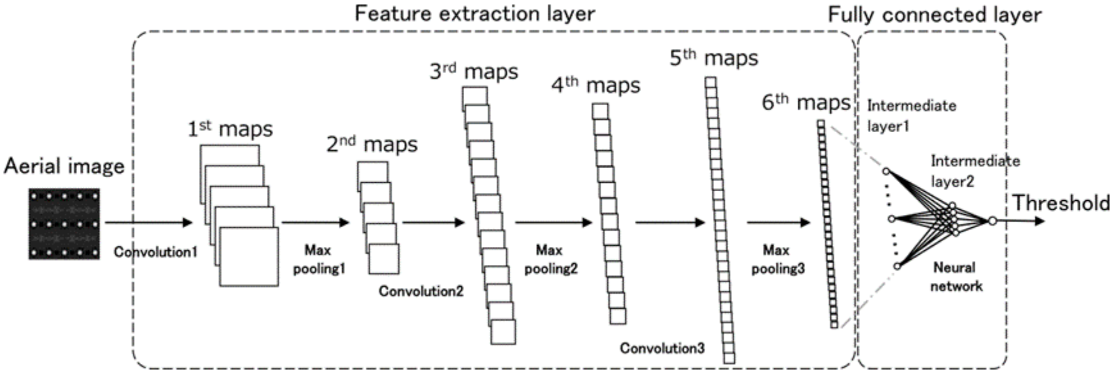


Figure 4. CNN architecture[14].

image is computed approximately with methods such as Hopkins method et.al. The chemistry involved in the resist model is complicated to compute from the first principle[16-18]. Sometimes, a simple threshold model is applied for the resist model and the threshold can be either constant or variable[19]. The etch model is more intractable due to the complicated physical- chemical process and multi-factorial control parameters involved such as plasma nature, chamber configuration et.al. Historically, variable etch bias (VEB) model is applied for the optimization purpose[20]. However, the approaches mentioned above may not be able to meet the requirement of the OPC techniques developing for the advanced technological nodes and more accurate and rigorous models are necessary while the nodes shrink. The machine learning based resist model and etch model turn out to be effective and becomes good candidates for future OPC application.

The general purpose of the ML based simulators is to obtain a general function approximator with the local geometric features as input and values of the height or threshold at the pixel level as the output. In principle, it can be done with the multilayer neural networks.

Seongbo Shim et. al. applied the full connected neural network to fit the resist model with the points sampled from the geometry of layout as the input and the resist height at center of the window as the output. The configuration of their model is shown in Figure 3. Youngchang Kim et. al. use the similar method to realize the prediction of etch bias[20].

Since the inventions of new architectures of neural networks emerge, more efficient and suitable approaches are fetched by the OPC community to improve the performance of the forward simulators. Yuki Watanabe et. al. use convolutional neural networks which are widely applied in the computer vision computation to estimate the resist pattern instead[14]. The architecture of their net is shown in Figure 4. Since sometimes, the rigorous simulation or experimental data are hard to obtain especially for new technical nodes, to obtain the trained model with the required accuracy with fewer data, Yibo Lin et. al. take the advantage of the transfer learning and active learning while they are trying to solve the same problem[22]. Later, the generative adversarial net is also introduced by the same group for simulating purpose[23].

3.2. Machine Learning Based Optimizers

Early days, the mask optimization relies on the empirical rules which usually depend on the geometry of the layout patterns. The contours of the layout patterns are decomposed into the edges and corners and the positions of their end points are varied and optimized according to the rules subtracted from the experimental facts[24]. Even though computationally efficient, the rule based optimization methods are not able to provide the required accuracy and fail to fulfill the requests of the advanced technological nodes. A more robust iterative optimization method based on the optical models, photoresist models et. al. is introduced. The basic idea is to change the positions of the end points of the edges and corners mentioned above and the simulated images on the wafer are obtained accordingly. The full optimization cycle is stopped once the target patterns and the simulated images match each other. Different kinds of error functions are applied to provide a quantitative estimation of the deviations between the target patterns and the simulated images. The Edge Displacement Error (EDE)[25], Edge Placement Error (EPE)[26] or Pixel-wise Error Summation[27] et. al. are usually calculated. The optimization process is usually computationally expensive due to the slow convergence of the iterations. While coarser models which are more computationally tractable are applied at the cost of the accuracy, the effort of the reduction of the iterations inspired the early application of machine learning techniques to the OPC regime and it remains as a main purpose of machine learning based OPC packages till nowadays.

The earlier attempt to obtain a better initial guess for the mask optimization process with the linear regression methods done by the researchers at University of California, Berkeley becomes an excellent start point in this track[28]. Taking advantage of the large dataset of the modified mask patterns after OPC by the commercial EDA packages, the authors estimate the expected fragment movement by the simplest linear statistical model provided an input target layout pattern. It provides a prototype of the basic ideas of the machine learning based OPC technologies within the framework of the supervised learning method. The realization of the workflow still requires the involvement of the advanced commercial EDA packages to generate the labeled data (the correct fragment movement given a specific mask pattern as the original input). As a result, the upper bond of the accuracy of this type of method is constrained by the correctness of the simulators and the efficiencies of the optimizers of the relevant commercial software. And the performance of such method is further compromised by the oversimplification of the mapping from the original input mask pattern to the predicted fragment movements by the application of a linear model. *However, the trained linear model serves as a coarse optimizer of the input mask pattern which provides the optimized mask pattern in a single run while it is fed by the input feature vectors representing the original mask patterns.* The authors successfully reduce the number of iterations of the traditional OPC workflow by the replacement of the original mask pattern with the statistically learned one as the initial condition of the subsequent optimization flow. *Another significant contribution of the authors is that they successfully introduce a representation method for the input mask layout which makes the further calculation they complete computationally feasible.* The discrete cosine transform (DCT) is applied to the input mask layout, and first hundreds of the DCT coefficients is collected in the Zig-Zag order (Shown in Figure 5.) as the input feature vectors of the linear model to be trained.

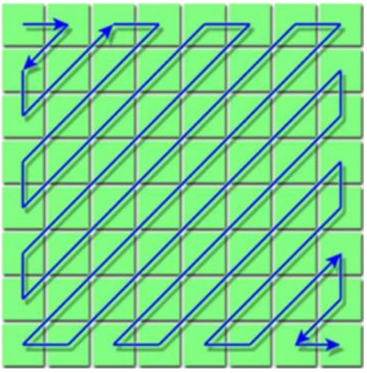


Figure 5. Zig–zag ordering of DCT coefficients[28].

The feature engineering accomplished this way serves as one of the mainstream techniques in the OPC community before more efficient and universal feature learning techniques fitting the requirement of the end to end learning such as the prevalent convolutional neural network (CNN) techniques are introduced from the deep learning community. The DCT is also applied by other researchers in the OPC community later in different ways including the variant form of the Fourier Transforms[29–32]. Even after the CNN et. al. deep learning techniques are introduced and the representation learning is realized automatically independent of the input data

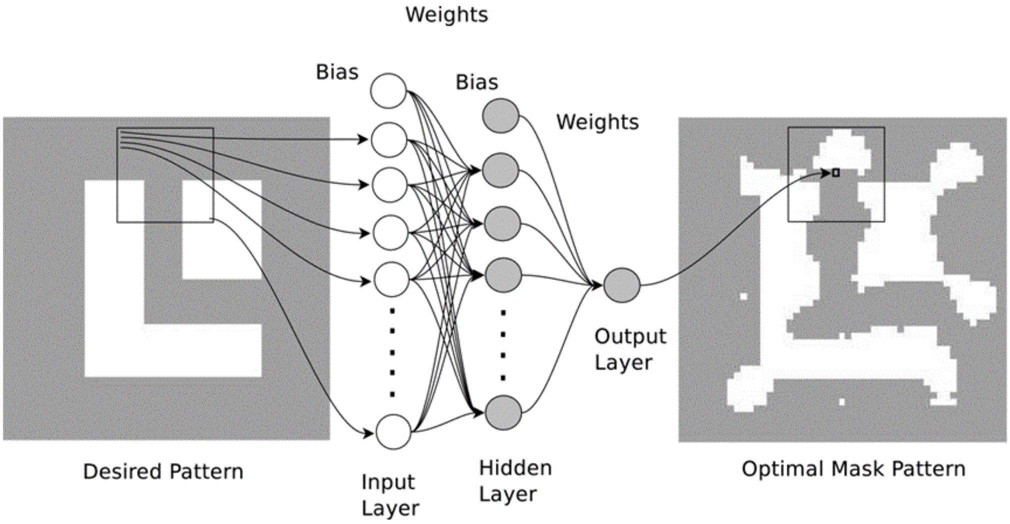


Figure 6. The schematics of the NN for OPC [36].

formulations, DCT is sometimes still used as the pre-processed data type as the neural network inputs[33].

The trained models after the supervised learning as the optimizer instead of the traditional iterative optimization circle are further improved mainly in two aspects: more complicated and accurate model instead of the linear statistical model are used to approximate the mapping between the input mask pattern and the optimized mask pattern (the optimization can be in the form of either motions of the specific edge fragment or the modified mask patterns as a whole.); Different feature engineering can be done or the representation learning within the scope of the deep learning can be applied to the mask pattern and the dimensional reduction can be realized in varied ways accordingly[34].

A direct improvement of the representation capability of the linear model has been done by Tetsuaki Matsunawa et. al.[35] by the application of the generalized linear mixed model instead to include the edge type effect. Considering the universal approximation property of the multilayer neural network, replacing the linear model with the typical multilayer neural network becomes another natural choice and has been done by Rui Luo[36]. The author considering the estimation of the binary value of the central pixel of the square modified mask pattern by the standard three layer neural network with the original pixel level binary mask pattern as the input instead of estimating the motion of the central fragment. To obtain the whole modified mask pattern, the author has to scan the three layer model over the original mask pattern. The schematics of the NN is shown in Figure 6. Such kind of scanning can be done naturally by the introduction of the convolutional neural networks and the three layer neural network above can actually be treated as the convolutional layer.

The contemporary convolutional neural networks (CNN) with varied architectures have been invented and widely applied to different scenes such as image segmentation, object recognition, image classification et. al.[37]. Basically, it is critical that the actual input of the prevalent CNNs is usually the tensor type data instead of the flatten one used in the Rui’s work, and the convolution layer/Kernel layer with the shared weight parameters slides across the input tensor. The pooling layers are usually applied to further reduce the dimensions of the features learned. After the invention of the training methods of the deep neural networks such as the backpropagation et al.[38], the CNNs emerges. The critical advantage of the deep CNNs is that they permit the representation learned from the multiple levels of the abstraction which are realized by the stacking of varied convolutional kernels and pooling layers. It avoids the necessity of the designing effort of feature engineering by human wisdom and enables the end to end training of models which can be widely applied. The CNNs are immediately fetched by the OPC community and relevant works have been done recently. Once we constrain our discussion within the mask pattern optimization or source optimization problems, the representation of the image patterns by the latent vectors and their decoding are naturally involved and can be directly linked to the encoder-decoder structures. For example, the convolutional autoencoder is trained to do the Source Mask Optimization by Ying Chen et. al.[39] to dramatically raise the speed of the

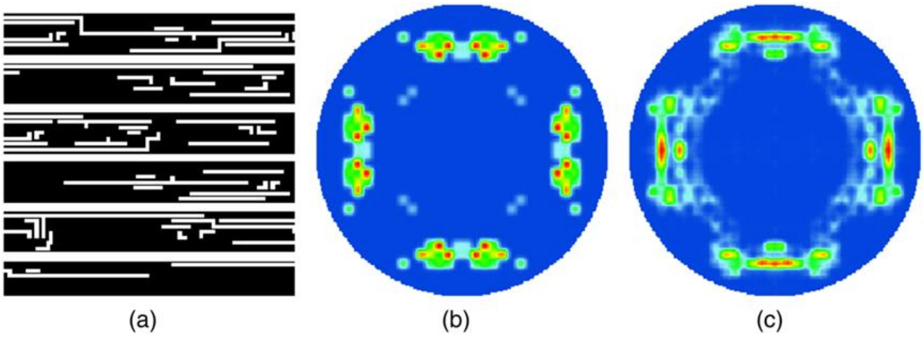


Figure 7. Illustrations of (a) a layout clip, (b) a model-based source, and (c) an autoencoder-based source [39].

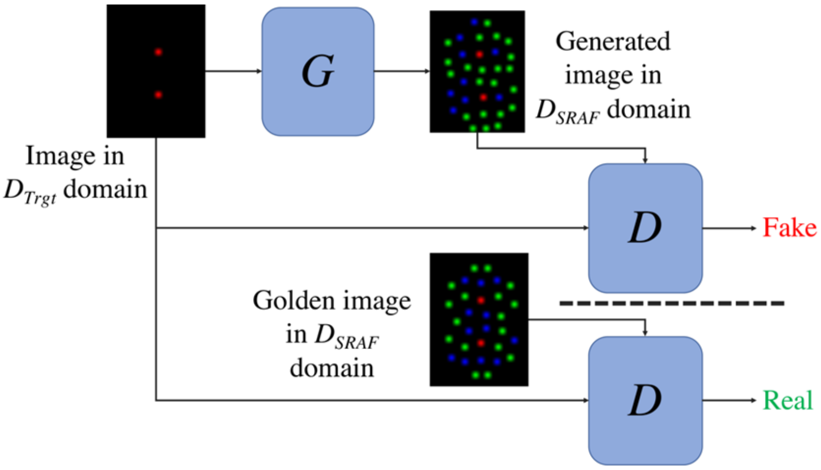

G
Generated image in D_{SRAF} domain
Image in D_{Trgt} domain
D
Fake
Golden image in D_{SRAF} domain
D
Real

Figure 8. An overview of the CGAN functionality[42].

optimization process by a factor of 10^5. Their model output is shown in Figure 7.

Similarly, the stacking convolutional architectures are also implemented by Haoyu Yang et al.[40] to form the generator and discriminator of the generative adversarial network (GAN)[41] when they succeed in realizing the mask optimization with the modified discriminator design. After the GAN converges, the generator can be used to calculate the optimized mask pattern of the original input one within 0.2s which is negligible compared with the traditional OPC methods. The convolutional autoencoders (CAE) are also applied in other regimes such as the insertion of the Sub Resolution Assist Features (SRAF) et. al.[42]. They can be trained as GAN shown in Figure 8.

Basically transformed into a image generation or translation problem[43, 44], the graphic generation of the modified mask pattern can be done by the mainstream computer vision techniques. Proper modifications made to the design of the specific architectures are necessary. Autoencoders can serve as the models or function approximators of the mapping between the input mask pattern and optimized mask pattern. The training process or the learning of the relevant parameters are finished in the supervise learning mode. In fact, the trained models as the optimizers are not necessarily functioned as the generators of the optimized mask or source patterns. They can also be easily applied as the classifiers for other OPC purposes. We are trying to separate these applications into different categories of OPC techniques although mathematically they are the same in the sense that they eventually act as function approximators providing the appropriate mappings minimizing the designed loss functions. The output can be either mask patterns, source patterns or the labels. We will leave these discussion to the next section where the pattern selection and hotspots detection *et. al.* are discussed.

3.3. Machine Learning Modified Workflow

As discussed by Peter De Bisschop[26], the whole OPC workflow strongly depends on the data collection and selection. The main point is: firstly, the feature structures among the billions on the VLSI chip should be selected to build the empirical models unless the physical process is clear enough to be simulated in the first principle way. The latter is rarely the case we confront in the realistic optical lithographic and etching processes. So the establishment and verification of the models as the simulators of the lithographic or etching processes require the data collection and selection even before the machine learning techniques are widely introduced into the OPC regime; secondly, after OPC process, the modified mask patterns or the source conditions should be verified both by the computational method (computational verification) and experimental method (on-wafer verification) before the masks are accepted for the production. As a result, feature pattern selection for the model calibration and on wafer verification et. al. become critical steps. The data sampling problem becomes important for an efficient and robust OPC workflow. The machine learning techniques can solve such kind of problems well. The basic idea is that we should be able to find a proper space defined with correct basis, in which the dimensional reduction of the original data set can be naturally realized. Or, the low dimension manifold in a high dimensional space is discovered and the sampling is done on the manifold only. Both methods can dramatically reduce the required number of sampling points and the cost of the time consuming and expensive computational or experimental verification processes. Dmitry Vengertsev et. al.[45] define a hybrid space formed by the direct sum of image parameter space and geometric sensitivity space and use a modified K means method to cluster the data within the hybrid space. As a typical unsupervised learning method, data clustering helps the selection of the representative patterns and serves as a kind of dimension reduction process. Instead of the K means method, the singular value decomposition (SVD) which can be treated as a form of the principle component analysis (PCA) can also be applied to the matrix representation of the layout patterns defined in the vector space manually constructed[46].

We already discuss the importance of the latent feature vector generation under the background of the machine learning based optimizer. It is also the foundation of the pattern selection we just discussed because the dimension reduction we mentioned is actually finished by the learning of a low dimensional representation of the original dataset. Now, the same thing goes with the hotspots detection. We need to identify the layout structures which can not be manufactured with the acceptable EPE et. al. under the current process conditions and carry out finer OPC for them. We are not able to carry out the forward simulation for all the structures on the chip due to the huge computational power that requires, or we just want a better solution[47]. We are neither satisfied with the traditional pattern match method[48, 49] because it can not predict the hotspot correctly when patterns not included in the library are met. Transforming such problems into the image classification problem[50] and solving it with the prevailing machine learning techniques then become interesting. *The basic idea is we learn the low dimensional feature vectoral representation of the layout patterns and use the classifier to distinguish the pattern with hotspots from the pattern without hotspots within certain region in the latent space formed by the learned feature vectors. You can also use them to do data clustering and realize the pattern feature selection.* The effectiveness of such kind of method strongly depends on the generalization capability of the machine learning model. It is not well understood when the learned model generalizes well especially when the deep learning techniques are applied. Even without the theoretical guarantee, these machine learning methods are applied in the hotspots detection widely and they are proven effective by the experimental facts. Matsunawa et. al.[51] use the human designed feature vectors to do the classification for the hotspots detection with Adaboost method. Taking advantage of the end to end training capability of deep CNNs, Moojoon Shin[52] et. al. apply different architectures of CNN binary classifier to fulfill the speed and accuracy requirement of hotspots detections. The probability of a pixel being classified as the hotspot is predicted by inputting the image centered at that pixel into the CNN. After scanning the whole layout, the probabilistic distribution of the hotspots at the pixel level is output as the final result. The schematics is shown in Figure 9.

Of course, even CNNs have the advantage in the sense that they automatically include the translational invariance and tend to learn the local information of image while encoding thus dramatically reduce the number of learnable parameters, the general fully connected deep neural network (DNN) can also be applied to carry out the hotspots detection task[53]. To improve the performance of the DNN hotspots detectors, different variants of DNN have been explored[33]. For example, inception mechanism is introduced by

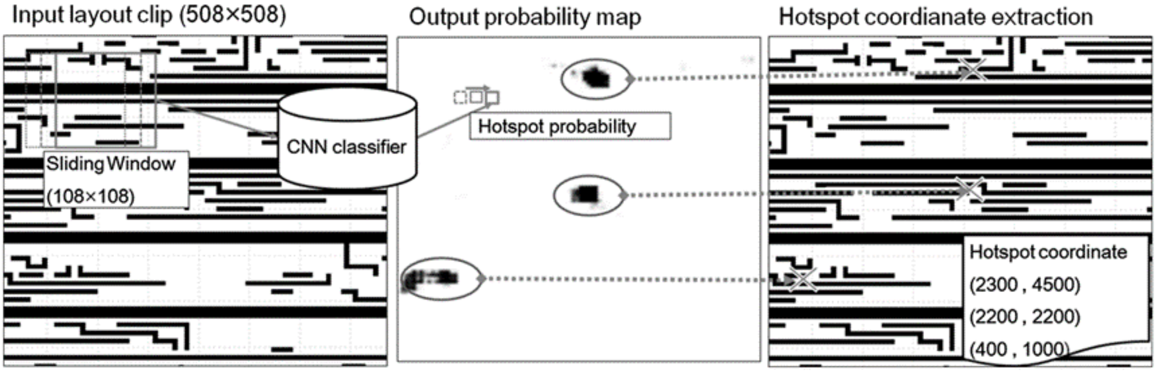


Figure 9. HS detection using sliding window scan and coordinate extraction[52].

Ran Chen et. al.[54]. Haoyu Yang et. al. modifies the CNN architecture and replace all the pooling layers with 3×3 convolution layers[55].

4. Conclusions

Machine learning techniques especially the deep learning method can dramatically improve the accuracy and computation speed of simulation and optimization process and the full chip level optimization techniques should become available and it will further change the whole workflow of current OPC technology[56].

Acknowledgments

This work is supported by National Science and Technology Major Project of China (2017ZX02315001-003, 2017ZX02101004-003), National Natural Science Foundation of China (61874002, 61804174), Beijing Natural Fund (4182021).

References

[1] J. W. Thackeray, “Stochastic exposure kinetics of extreme ultraviolet photoresists: simulation study,” *J. Micro/Nanolithography, MEMS, MOEMS* **10**(3), 033019 (2011).

[2] A. Erdmann, T. Fühner, F. Shao, and P. Evanschitzky, “Lithography simulation: modeling techniques and selected applications,” *Model. Asp. Opt. Metrol. II* **7390**, 739002 (2009).

[3] A. K. Wong and A. R. Neureuther, “Mask Topography Effects in Projection Printing of Phase-Shifting Masks,” *IEEE Trans. Electron Devices* **41**(6), 895–902 (1994).

[4] P. Evanschitzky and A. Erdmann, “Three dimensional EUV simulations: a new mask near field and imaging simulation system,” *25th Annu. BACUS Symp. Photomask Technol.* **5992**, 59925B (2005).

[5] K. Adam and A. R. Neureuther, “Domain decomposition methods for the rapid electromagnetic simulation of photomask scattering,” *J. Microlithogr. Microfabr. Microsystems* **1**(3), 253–269 (2002).

[6] J. Byers, J. Petersen, and J. Sturtevant, “Calibration of Chemically Amplified Resist Models,” *Proc. SPIE* **2724**, 156–162 (1996).

[7] A. Erdmann, G. Citarella, P. Evanschitzky, H. Schermer, V. Philipsen, and P. De Bisschop, “Validity of the Hopkins approximation in simulations of hyper-NA (NA>1) line-space structures for an attenuated PSM mask,” *Opt. Microlithogr.* XIX **6154**, 61540G (2006).

[8] S. M. Kim, S. J. Kim, C. J. Bang, Y. M. Ham, and K. H. Baik, “Optimization of dipole off-axis illumination by 1st-order efficiency method for sub-120 nm node with KrF lithography,” *Japanese J. Appl. Physics, Part 1 Regul. Pap. Short Notes Rev. Pap.* **39**(12 B), 6777–6780 (2000).

[9] S. Suh, Y. Kang, I. Kim, S. Woo, H. Cho, and J. Moon, “Pattern type specific modeling and correction methodology at high NA and off-axis illumination,” *25th Annu. BACUS Symp. Photomask Technol.* 5992(2005), 599220 (2005).

[10] H. Aoyama, Y. Mizuno, N. Hirayanagi, N. Kita, R. Matsui, H. Izumi, K. Tajima, J. Siebert, W. Demmerle, and T. Matsuyama, “Impact of realistic source shape and fiexibility on source mask optimization,” *J. Micro/Nanolithography, MEMS, MOEMS* **13**(1), 011005 (2014).

[11] N. Jia and E. Y. Lam, “Pixelated source mask optimization for process robustness in optical lithography,” *Opt. Express* **19**(20), 19384 (2011).

[12] R. Socha, X. Shi, and D. LeHoty, “Simultaneous source mask optimization (SMO),” *Photomask Next-Generation Lithogr. Mask Technol.* XII **5853**, 180 (2005).

[13] A. K. Wong, Resolution Enhancement Techniques in Optical Lithography, SPIE press (2001)

[14] Y. Watanabe, T. Kimura, T. Matsunawa, and S. Nojima, “Accurate Lithography Simulation Model based on Convolutional Neural Networks,” *Opt. Microlithogr.* XXX **10147**, 101470K (2017).

[15] A. K. Wong, *Optical Imaging in Projection Microlithography*, SPIE press (2005).

[16] C.M. Garza, C.R. Szmanda, and R.L. Fischer Jr., “Resist dissolution kinetics and submicron process control” *Proceedings of SPIE: Advances in Resist Technology and Processing V* **920**, 321–338. (1988)

[17] K. Itoh, K. Yamanaka, H. Nozue, and K. Kasama, “Dissolution kinetics of high resolution novolac resists”

Proceedings of SPIE: Advances in Resist Technology and Processing VIII **1466**, 485–496.(1991)

[18] C.A. Mack, M.J. Maslow, R. Carpio, and A. Sekiguchi, New model for the effect of developer temperature on photoresist dissolution, *Proceedings of SPIE: Advances in Resist Technology and Processing XV* **3333**, 1218–1231. (1998)

[19] Y. Granik, N. B. Cobb, and T. Do, "Universal process modeling with VTRE for OPC," *Opt. Microlithogr.* XV **4691**(2002), 377 (2002).

[20] Y. Kim, S. Jung, D. Kwak, V. Liubich, and G. Fenger, "Predictable etch model using machine learning," Opt. Microlithogr. XXXII 10961, 1096106 (2019).

[21] S. Shim, S. Choi, and Y. Shin, "Machine learning-based 3D resist model," Opt. Microlithogr. XXX 10147(March 2017), 101471D (2017).

[22] Y. Lin, M. Li, Y. Watanabe, T. Kimura, T. Matsunawa, S. Nojima, and D. Z. Pan, "Data Efficient Lithography Modeling with Transfer Learning and Active Data Selection," *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **38**(10), 1900–1913 (2019).

[23] W. Ye, M. B. Alawieh, Y. Lin, and D. Z. Pan, "LithoGAN: End-to-end lithography modeling with generative adversarial networks," *Proc. - Des. Autom. Conf.* 2019 (2019).

[24] J. Park, C. Park, S. Rhie, Y. Kim, M. Yoo, J. Kong, H. Kim, and S. Yoo, "An Efficient Rule based OPC Approach Using a DRC tool for 0.18 um ASIC," *Proc. IEEE First Int. Symp. Qual. Electron. Des.*, 81–85 (2000).

[25] W. Ye, M. B. Alawieh, Y. Lin, and D. Z. Pan, "LithoGAN: End-to-End Lithography Modeling with Generative Adversarial Networks," *Proc. DAC* **107**, 1–6 (2019).

[26] P. De Bisschop, "Optical proximity correction : A cross road of data flows Characteristics in Extreme Ultraviolet Lithography," *Jpn. J. Appl. Phys.* **55**, 06GA01 (2016).

[27] Y. Shen, N. Jia, N. Wong, and E. Y. Lam, "Robust level-set-based inverse lithography," *Opt. Express* **19**(6), 5511 (2011).

[28] A. Gu and A. Zakhor, "Optical proximity correction with linear regression," *IEEE Trans. Semicond. Manuf.* **21**(2), 263–271 (2008).

[29] S. Shim and Y. Shin, "Topology-oriented pattern extraction and classification for synthesizing lithography test patterns," *J. Micro/Nanolithography, MEMS, MOEMS* **14**(1), 013503 (2015).

[30] T. Matsunawa, B. Yu, and D. Z. Pan, "Laplacian eigenmaps- and Bayesian clustering-based layout pattern sampling and its applications to hotspot detection and optical proximity correction," *J. Micro/Nanolithography, MEMS, MOEMS* **15**(4), 043504 (2016).

[31] S. Shim, W. Chung, and Y. Shin, "Synthesis of lithography test patterns through topology-oriented pattern extraction and classification," *Des. Co-optimization Manuf.* VIII **9053**, 905305 (2014).

[32] W. Zhang, X. Li, S. Saxena, A. Strojwas, and R. Rutenbar, "Automatic clustering of wafer spatial signatures," *Proc. - Des. Autom. Conf.*, 1–6 (2013).

[33] H. Yang, J. Su, Y. Zou, Y. Ma, B. Yu, and E. F. Y. Young, "Layout Hotspot Detection with Feature Tensor Generation and Deep Biased Learning," *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **38**(6), 1175–1187 (2019).

[34] G. Hinton and R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science* (80-.). 313(July), 504–507 (2006).

[35] T. Matsunawa, B. Yu, and D. Z. Pan, "Optical proximity correction with hierarchical Bayes model," *Opt. Microlithogr.* XXVIII **9426**, 94260X (2015).

[36] R. Luo, "Optical proximity correction using a multilayer perceptron neural network," *J. Opt.* (United Kingdom) **15**(7) (2013).

[37] W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," *Neural Comput.* **29**, 2352–2449 (2017).

[38] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).

[39] Y. Chen, Y. Lin, L. Dong, T. Gai, R. Chen, Y. Su, Y. Wei, and D. Z. Pan, "SoulNet: ultrafast optical source optimization utilizing generative neural networks for advanced lithography," *J. Micro/Nanolithography, MEMS, MOEMS* **18**(04), 1 (2019).

[40] H. Yang, S. Li, Z. Deng, Y. Ma, B. Yu, and E. F. Y. Young, "GAN-OPC: Mask Optimization with Lithography-guided Generative Adversarial Nets," *IEEE Trans. Comput. Des. Integr. Circuits Syst.* (2019).

[41] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Adv. Neural Inf. Process. Syst.* **3**(January), 2672–2680 (2014).

[42] M. B. Alawieh, Y. Lin, Z. Zhang, M. Li, Q. Huang, and D. Z. Pan, "GAN-SRAF: Sub-Resolution Assist Feature Generation using Generative Adversarial Networks," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*(i), 1–6 (2020).

[43] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," 1–7 (2014). http://arxiv.org/abs/1411.1784.

[44] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 1063–6919 (2017).

[45] D. Vengertsev, K. Kim, S.-H. Yang, S. Shim, S. Moon, A. Shamsuarov, S. Lee, S.-W. Choi, J. Choi, and H.-K. Kang, "The new test pattern selection method for OPC model calibration, based on the process of clustering in a hybrid space," *Photomask Technol.* **2012** 8522, 85221A (2012).

[46] Y. Sun, Y. M. Foong, Y. Wang, J. Cheng, D. Zhang, S. Gao, N. Chen, B. Il Choi, A. J. Bruguier, M. Feng, J. Qiu, S. Hunsche, L. Liu, and W. Shao, "Optimizing OPC data sampling based on 'orthogonal vector space,'" *Opt. Microlithogr.* XXIV **7973**, 79732K (2011).

[47] J. Kim and M. Fan, "Hotspot detection on post-OPC layout using full-chip simulation-based verification tool: a case study with aerial image simulation," 23rd Annu. BACUS Symp. *Photomask Technol.* **5256**, 919 (2003).

[48] W. Y. Wen, J. C. Li, S. Y. Lin, J. Y. Chen, and S. C. Chang, "A fuzzy-matching model with grid reduction for lithography hotspot detection," *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **33**(11), 1671–1680 (2014).

[49] Y. T. Yu, Y. C. Chan, S. Sinha, I. H. R. Jiang, and C. Chiang, "Accurate process-hotspot detection using critical design rule extraction," *Proc. - Des. Autom. Conf.*, 1167–1172 (2012).

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM 60(6), 84–90 (2017).

[51] T. Matsunawa, J.-R. Gao, B. Yu, and D. Z. Pan, "A new lithography hotspot detection framework based on AdaBoost classifier and simplified feature extraction," *Des. Co-optimization Manuf.* IX **9427**, 94270S (2015).

[52] M. Shin and J.-H. Lee, “Accurate lithography hotspot detection using deep convolutional neural networks,” *J. Micro/Nanolithography, MEMS, MOEMS* **15**(4), 043507 (2016).

[53] T. Matsunawa, S. Nojima, and T. Kotani, “Automatic layout feature extraction for lithography hotspot detection based on deep neural network,” *Des. Co-optimization Manuf.* X **9781**, 97810H (2016).

[54] R. Chen, W. Zhong, H. Yang, H. Geng, X. Zeng, and B. Yu, “Faster region-based hotspot detection,” *Proc. - Des. Autom. Conf.*, 0–5 (2019).

[55] H. Yang, Y. Lin, B. Yu, and E. F. Y. Young, “Lithography hotspot detection: From shallow to deep learning,” Int. Syst. Chip Conf. 2017-Septe, 233–238 (2017).

[56] I. Torunoglu, A. Karakas, E. Elsen, C. Andrus, B. Bremen, B. Dimitrov, and J. Ungar, “A GPU-based full-chip inverse lithography solution for random patterns,” Des. Manuf. through Des. Integr. IV 7641(April 2010), 764115 (2010).

[57]

Photography & Biography

Pengpeng Yuan received his BS degree in Integrated circuits engineering from Tsinghua University, Beijing, China. He is currently working toward the Ph.D. degree in Institute of Microelectronics of the Chinese Academy of Science, Beijing, China. His research interests include computational imaging, lithographic resolution enhancement based on deep learning.

Taian Fan is a research assistant at IMECAS (Institution of Microelectronics Chinese Academy of Science). He received his BS in Electrical Engineering from the Beijing Jiaotong University in 2013, and his Ms. degree in Electrical Engineering from the University of Vermont in 2016. His current research interests include Computational Lithography, Computational Electrodynamics and Design/Simulation automation.

Yaobin Feng is senior director in charge of Lithography in YMTC, he joined YMTC in charge of Lithography technology development since 3D NAND project kicked off in 2015. He graduated from Shanghai Jiaotong University as MBA and Southeast University as Bachelor of Automation. Before joining YMTC, He worked in Micron, UMC and ASMC since 2003. He published many papers in computational lithography, lithography process and metrology.

Peng Xu received his BS degree in Applied Physics from Harbin Institute of Technology, China and PhD degree in Physics from College of William and Mary, USA. He was a postdoctoral scholar in Institute of Physics, Chinese Academy of Science. He is currently an associate professor at IMECAS. His research interests include computational lithography, near field optical imaging and optical spectroscopy.

Yayi Wei is a “Ten Thousand Talents Plan” professor at IMECAS, and he also serves as the director of the Computational Lithography R&D Center and the Open Laboratory of Zhongguancun, focusing on the computational lithography research for the advanced technology node. Prof. Wei received his Ph.D. from the Max Planck Institute for Solid State Research/Stuttgart University under the guidance of the Nobel Prize winner Klaus von Klitzing. Prior to IMECAS, he worked in many prestigious institutions and enterprises including the Department of Energy's Oak Ridge National Laboratory, Infineon New York R&D Center of the U.S.A. and GLOBALFOUNDRIES New York R&D Center of the U.S.A. Prof. Wei has long been engaged in the research and development of semiconductor devices, materials and processes in the semiconductor lithography field. He led or participated in various projects from the 180nm to 10nm technology node.

J. Microelectron. Manuf. **3**, 20030409 (2020)
doi: 10.33079/jomm.20030409

Enabling Variability-Aware Design-Technology Co-Optimization for Advanced Memory Technologies

Salvatore M. Amoroso[1], Plamen Asenov[1], Jaehyun Lee[1], Nara Kim[2], Ko-Hsin Lee[3], Yaohua Tan[4], Yong-Seog Oh[4], Lee Smith[4], Xi-Wei Lin[4, *] and Victor Moroz[4]

[1] *Synopsys Europe, Ltd., Glasgow, G3 8HB, UK*
[2] *Synopsys Korea, Pankyoyeokro 235, Gyeonggi-do13494 South Korea*
[3] *Synopsys Taiwan Co., Ltd., Chupei 302, Taiwan.*
[4] *Synopsys, Inc., Mountain View, CA 94043 USA*

Abstract: This paper presents a TCAD-based methodology to enable Design-Technology Co-Optimization (DTCO) of advanced semiconductor memories. After reviewing the DTCO approach to semiconductor devices scaling, we introduce a multi-stage simulation flow to study the device-to-circuit performance of advanced memory technologies in presence of statistical and process variability. We present a DRAM example to highlight the DTCO enablement for both memory and periphery. Our analysis demonstrates how the evaluation of different possible technology improvements and design combinations can be carried out to maximize the benefits of continuous technology scaling for a given set of manufacturing equipment.

Keywords: DTCO, Statistical Variability, Process Variability, Semiconductor Memories, DRAM, CMOS, Scaling.

1. Introduction

The pace of the technology roadmap for semiconductor was conventionally marked by scaling of the patterning pitches, with the main goal to halve the cost per transistor at each subsequent technology node. A certain level of uncertainty affecting the time-to-market of a technology node is intrinsic in this scaling approach. Today, the semiconductor industry is facing a paradigm shift, with scaling now being driven by annual technology releases for both memory and logic. This new approach is driven by schedule to deliver the best possible combination of technology improvements within a year. In order to support this endeavour, the semiconductor industry has adopted a Design-Technology Co-Optimization (DTCO) methodology, which requires fundamental figures of merit, namely Power-Performance-Area (PPA) or its variant Power-Performance-Area-Cost (PPAC), to be evaluated and optimized across a set of different possible technology improvements to maximize the gain brought by each annual technology update [1]–[6]. Furthermore, memory manufacturing has to deal with specific set of challenges, which are ruled by parametric yield and process window optimization for both periphery and the memory cell [7]–[10].

In this paper we will use a DRAM example to highlight the DTCO enablement for both memory and periphery. DRAM represents a well-suited test-bed because the continuing efforts in its processing technology have enabled dramatic feature-size reduction and unprecedented levels of integration [11]–[14], but also increased the severity of parasitic effects [15]. In particular, during the design cycle, attention has to be put on the DRAM cell transistor leakage current, which dictates DRAM refresh time (tREF) and, in turn, affects manufacturing yields. It is of utmost importance to highlight that the DTCO methodology cannot be focused to the average circuit behaviour. Indeed, the ultimate failure in yield is governed by the leakage current of extreme-tail cells (<10^{-6} probability). These cells may exhibit a few orders of magnitude higher leakage than the nominal cell, with a statistical distribution that is influenced by both process (e.g. geometry, doping profiles) and intrinsic statistical variability (e.g. random discrete dopants, random traps). Although innovative characterization techniques have been proposed to experimentally evaluate the DRAM cell transistor leakage current distributions [16], it becomes also essential to have available modelling platforms that enable a fully variability-aware Design-Technology Co-Optimization (DTCO) of DRAM circuits to evaluate and optimize DRAM yields in the presence of process and statistical

[*] Address all correspondence to Xi-Wei Lin, E-mail: xiwei@synopsys.com

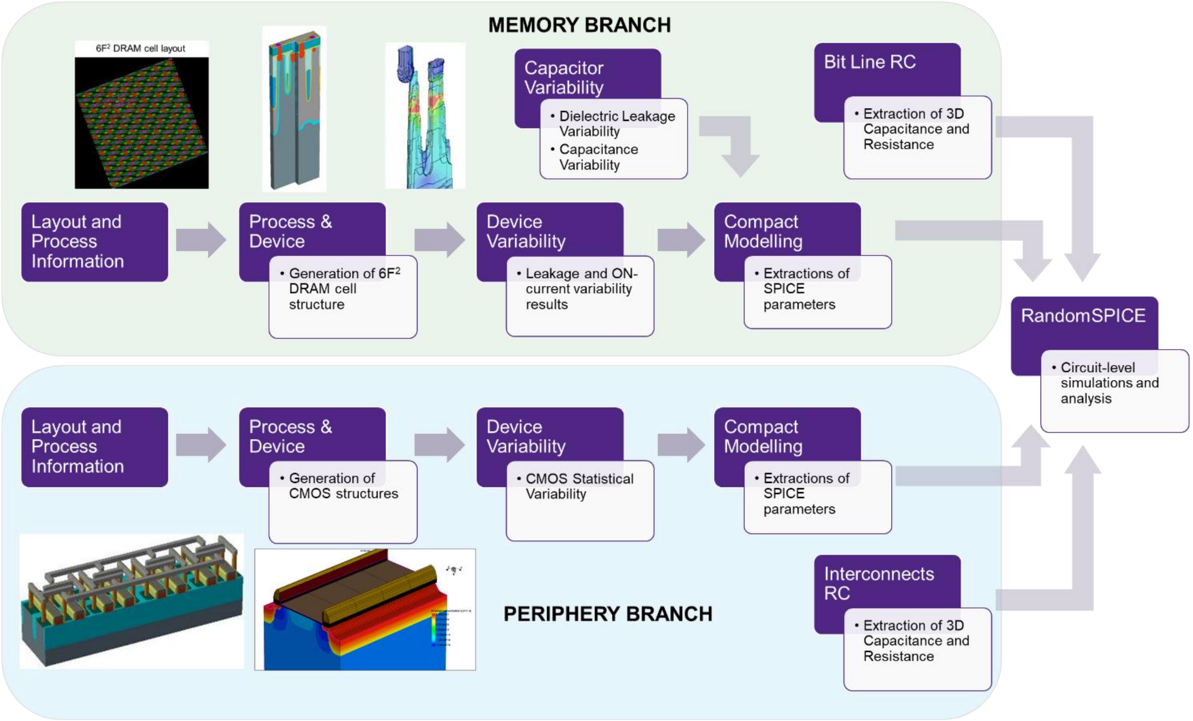

MEMORY BRANCH
6F² DRAM cell layout
Capacitor Variability
• Dielectric Leakage Variability
• Capacitance Variability
Bit Line RC
• Extraction of 3D Capacitance and Resistance
Layout and Process Information
Process & Device
• Generation of 6F² DRAM cell structure
Device Variability
• Leakage and ON-current variability results
Compact Modelling
• Extractions of SPICE parameters
RandomSPICE
• Circuit-level simulations and analysis
Layout and Process Information
Process & Device
• Generation of CMOS structures
Device Variability
• CMOS Statistical Variability
Compact Modelling
• Extractions of SPICE parameters
Interconnects RC
• Extraction of 3D Capacitance and Resistance
PERIPHERY BRANCH

Figure 1. Simulation-based DTCO methodology for the DRAM refresh time optimization in presence of statistical and process variability.

variability with reduced requirements on costly and slow silicon manufacturing cycles.

The remainder of the paper is organized as following: Section 2 introduces our simulation-based DTCO methodology; Section 3 presents the DTCO simulation results for the memory part, including variability and reliability issues affecting write and retention operations; Section 4 presents the DTCO simulation results for the periphery circuit (Sense Amplifier) including variability and interconnect parasistics analysis affecting the sensing operation; finally, Section 5 will summarize the results and draw the conclusions.

2. Simulation-based DTCO Methodology

In this paper we present a DTCO modelling approach enabling the optimization of memory and periphery performance for a DRAM array. The methodology includes the early injection of statistical metrics into the design/optimization cycle.

This multi-stage simulation flow, which allows accurate and extensive exploration of the design space by taking into account both memory and periphery performance figures of merit and their statistical behavior, consists of two branches (Figure 1): memory branch and periphery branch.

The memory branch (indicated with "M") targets the study and optimization of write and retention variability and it features the following steps: (i-M) accurate process structure generation for the memory cells by means of Process Explorer (layout to 3D structure) [17] and Sentaurus Process [18] to capture process and doping profile variations, (ii-M) accurate device simulation of the nominal transistors by means of Sentaurus Device [19], (iii-M) statistical simulation of leakage through capacitor dielectrics by means of the Kinetic Monte Carlo (KMC) engine of Sentaurus Device [19]; (iv-M) Garand VE [20] for the physics-based variability simulation of trap-assisted leakage current in presence of random discrete dopants (RDD), (v-M) Mystic [21] to extract statistical compact models; (vi-M) Raphael FX [22] to extract parasitic RC components, including bitline capacitance and resistance for a given layout.

The periphery branch (indicated with "P") targets the study and optimization of the sensing operation and it features the following steps: (i-P) accurate process structure generation for the CMOS part by means of Process Explorer (layout to 3D structure) and Sentaurus Process [17],[18] to capture process and doping profile variations, (ii-P) accurate device simulation of the nominal transistors by means of Sentaurus Device [19], (iii-P) Garand VE [20] for the physics-based variability simulation of CMOS transistors in presence of RDD, line edge roughness (LER), metal gate granularity (MGG) etc.

Table 1. Variability components affecting the DRAM refresh time addressed by our DTCO flow.

Variability Component	**Flow Branch**	**Simulation Tool**
DRAM Cell Process Variations	Memory	Process Explorer, S-Process
Storage Capacitor Write Variations	Memory	S-Device, Garand VE
Storage Capacitor Leakage Variations	Memory	S-Device KMC
DRAM Transistor Leakage Variations	Memory	Garand VE
DRAM Disturbs Variations (not included in this work, see ref [26])	Memory	S-Device
Cell Array RC Extraction	Memory	Raphael FX
Sense Amplifier Process	Periphery	Process Explorer, S-Process
Local Transistors Mismatch	Periphery	Garand VE
Interconnects RC extraction	Periphery	Raphael FX
Line-to-line Dielectric Reliability (not included in this work, see ref [27])	Memory/Periphery	S-Device KMC
Bitline and wordline profile variations (not included in this work)	Memory/Periphery	S-Litho, Proteus

(iv-P) Mystic [21] to extract statistical compact models; (vi-P) Raphael FX [22] to extract interconnects resistances and capacitances (RC).

The two branches are then merged together for a statistical SPICE simulation analysis including memory, periphery and parasitic components, which we perform by means of the Monte Carlo circuit generator RandomSpice [23] and HSPICE [24]. Table 1 summarizes the variability components affecting the refresh time of a DRAM cell, which are addressed by our DTCO flow. In this work we are neglecting variations associated with the reliability of the DRAM transistors (statistical Row-Hammer [26]) and interconnects (statistical dielectric leakage/breakdown [27]). Furthermore, this DTCO analysis could be extended by considering the bitline/wordline shape variations: indeed Optical Proximity Correction (OPC) simulation could be employed to generate geometrical contours that represent wide (best R worst C) and narrow (worst R best C) bitline/wordline, therefore evaluating the performance of these variation corners.

3. Memory DTCO Analysis

The goal of the simulation-based DTCO flow shown in Figure 1 is to achieve the simulation based estimation and optimization of the DRAM refresh time (tREF) and, in turn, DRAM yield, in presence of process and statistical variability and for a given set of manufacturing assumptions. In this section, we will address the issues limiting tREF at the memory array level, whilst in Section 4 we will focus on the CMOS periphery limitations (Table 1).

3.1. DRAM Transistor – Process and Statistical Variability

The Synopsys TCAD platform [17]–[23] is used for the generation and simulation of the 3D DRAM array. The DRAM structures are constructed by means of Process Explorer [17] starting from a 6F^2 tilted-cell layout representative of a 2z nm technology node (Figure 2). A single cell and two adjacent neighbors are then cut-out to perform accurate doping implantation and device simulation by means of S-Process [18] and S-Device [19], respectively. Different process conditions are simulated by changing WLetch (WL recess etch) and Dose (roll-off) parameters by +/- 20% (Figure 2) to generate a range of structures corresponding to different process conditions, or process variations. The cell transistor, consisting of a saddle-fin featuring buried metal WL and shared common BL (Table 2), is then re-meshed to enable the statistical simulation of ON and leakage currents by means of the drift-diffusion variability engine Garand VE [20].

It has been previously shown that discrete doping can play a fundamental role in determining the stochastic dispersion of both drive current and leakage current in transistors. In this work, we consider the trap-assisted band-to-band tunneling (TAT) as the dominant contribution to the transistor leakage. The experimental results, in fact, clearly show that the transistor leakage current is a function of the number of defects in silicon, their energy level in the bandgap, and the electric field [6]. The trap-

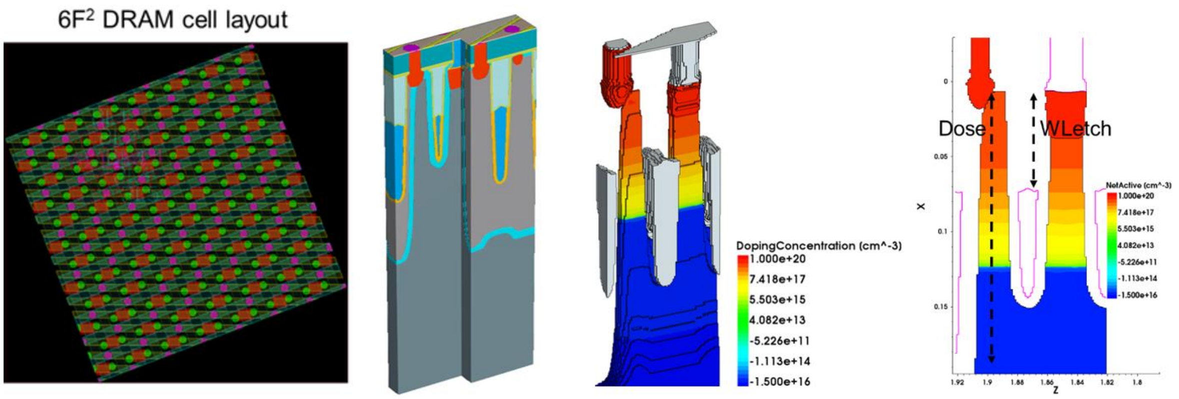


Figure 2. Layout to Process and Device simulation. Process variability is accounted for by varying the implantation dose and the gate height parameters by +/-20% with respect to the nominal process.

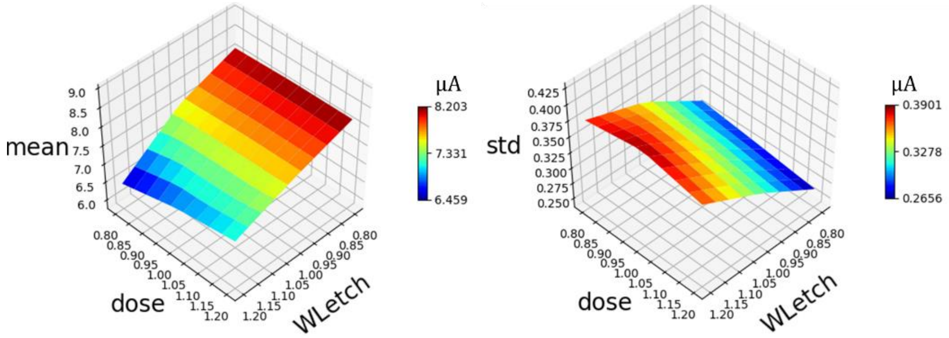


Figure 3. ON current average (left) and variability (right) performance across the space of process variations.

assisted contribution is modelled through an enhancement of the trap capture cross-section in the conventional Shockley-Read-Hall (SRH) generation term. The enhancement can either be computed by Hurkx-like local models or by non-local tunneling path approaches. For each process corner, Garand VE simulates hundreds of statistical instances. Each instance features a different configuration of random discrete dopants (RDD) and thousands of single-trap positions are evaluated to gather the TAT leakage statistics. Once the single-trap leakage statistics are obtained, any other statistics due to an arbitrary trap density can then be obtained at SPICE level by convolution of the single-trap cumulative distribution functions (as detailed in [28]).

Table 2. DRAM Transistor nominal dimensions and electrical parameters.

Critical Dimensions	
WLetch	60nm
Peak Dose	2e19cm^{-3}
Technology node	2z nm
Electrical Parameters	
V(core)	1.0V
V(bulk)	-0.8V
V(bbw)	-0.2V

Figure 3 shows the results of the Garand VE analysis performed to evaluate the impact of RDD on the ON-current for the DRAM cell, across the WLetch and Dose process variations space. A 10% variation in the mean ON-current can be observed, whilst the ON-current standard deviation varies from 3% to 6% of the nominal ON-current value. These variations can be understood by considering that the combination of WLetch and Dose define the gate to source/drain overlap. With a high WLetch, there is significant underlap, leading to low ON-current and high variability.

To evaluate the leakage variability, we have performed 200 Garand VE simulations for each process corner. For each RDD configuration, the single-trap TAT leakage is simulated by sweeping the trap position across the drain (storage node contact) pillar region with a 0.5nm spacing, leading to ~70,000 trap evaluations per each RDD configuration (14,000,000 trap configurations for each simulated process condition). Figure 4 shows the leakage complementary cumulative distribution, highlighting that the interaction between discrete traps and random dopants leads to extended exponential-like tails. Moreover, both average and tail behavior strongly depend on the process

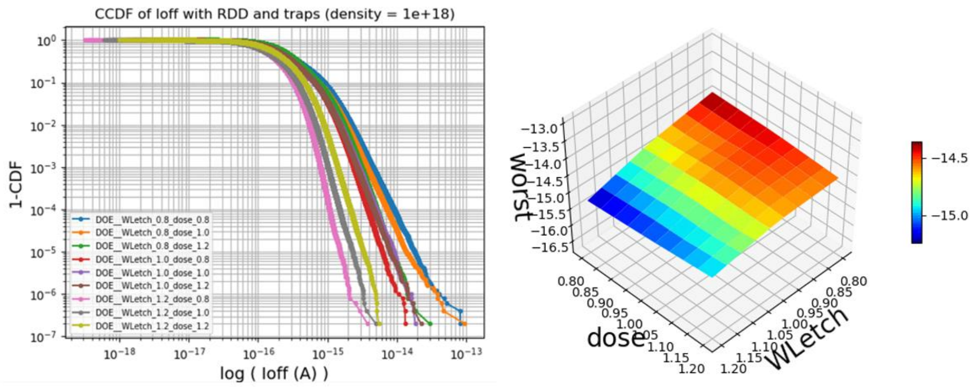


Figure 4. Leakage complementary cumulative distribution for different process corners (left); the worst leakage value is plotted across the space of process variation (right) as measure of the leakage variability.

variations. It is important to note that the variability of ON-current is anti-correlated to the variability of leakage. Therefore, the best process corner that minimizes ON-current variability is also be the worst corner that maximizes leakage variability. This imposes a trade-off between ON-current and leakage performance and, in turn, between DRAM write time (tWR) and tREF performance.

Once the statistical TCAD results are obtained across the space of process variations, compact models can be extracted by means of a response surface methodology in Mystic [21], as detailed and validated in [28]. It is worth remarking that the leakage due to many random traps can be obtained analytically by self-convolution of the single-trap statistics.

3.2. DRAM Capacitor Dielectric Leakage – Statistical Variability

DRAM capacitors utilize high-k dielectrics to maximize capacitance for a given technology node. Defects in high-k materials may cause undesirable leakage currents due to trap assisted tunneling. The leakage currents in the capacitors in a memory device have been one of the bottlenecks for further scaling down. Therefore, a systematic way of modeling and understanding the trap assisted tunneling transport mechanisms is required to support further downscaling.

To calculate the leakage current for a metal-insulator-metal structure, we have developed a stochastic reliability simulator, Sentaurus Device KMC [19], based on the kinetic Monte-Carlo method. The simulator randomly distributes discrete defects in insulator regions of a 3D capacitor structure. These discrete defects act as traps of carriers in an insulator that can affect device reliability. To simulate the electron transport via the traps, the electron hopping event rates are calculated with various physical models [29], including direct tunneling, Fowler-Nordheim (FN) tunneling, inelastic multi-phonon trap-to-trap and trap-to-electrode tunneling [30], and Poole-Frenkel (PF) emission [31]. The direct tunneling and FN tunneling are leakage currents without traps; they are determined by the intrinsic insulator properties. With the traps in an insulator, the inelastic multi-phonon processes dominate the tunneling current. These processes involve the emission and absorption of multiple phonons. In the PF emission, the localized electron in a trap is thermally excited to the conduction band of an insulator. Furthermore, the potential energy distribution is calculated by solving the Poisson equation with the image charge barrier lowering near electrodes as well as the short-ranged trap potentials.

With the KMC method, all possible electron transport events are considered as stochastic process [32]. The steady state current I_k is calculated by counting the net electrons at the electrode ΔN_k within Δt by I_k=(qΔN_k)/Δt, when the stochastic process reaches steady states.

Figure 5 shows the trap assisted tunneling current as a function of the electric field in a HfO_2 capacitor. The thickness of the HfO_2 layer is 5nm, and the outer diameter of the cylinder is 60nm. The electrodes are TiN. The leakage currents are compared according to the solid states of the insulator, i.e., monocrystalline, amorphous, and polycrystalline HfO_2. For the monocrystalline and amorphous HfO_2, the traps are randomly distributed in the bulks where the trap concentrations are $2×10^{19}$ cm^{-3} and the trap locations are identical for both structures. For the polycrystalline HfO_2, the same number of traps are distributed only on the grain boundaries, which result in smaller trap-to-trap

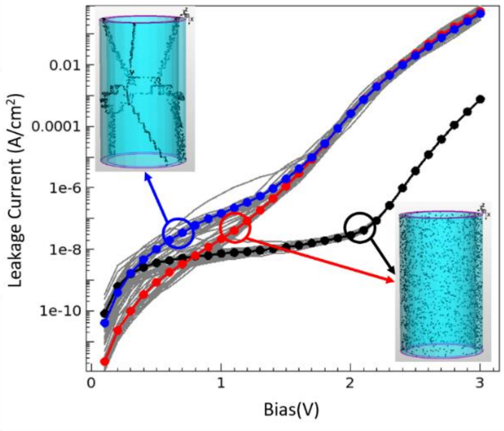


Figure 5. Leakage current in cylinder capacitors. Black line: Averaged current in the crystalline insulator, traps are randomly distributed in the bulk with the same trap energy of 1.8eV. Red line: Averaged current in the amorphous insulator, traps are randomly distributed in the bulk. Blue line: Averaged current in the polycrystalline insulator, traps are randomly distributed only on grain boundaries.

distances in the polycrystalline HfO_2. For the crystalline HfO_2, a constant trap level, 1.8 eV is used for all traps. In amorphous and polycrystalline HfO_2, the trap levels are randomly defined with the Gaussian distribution of the average 1.8 eV and the standard deviation 0.5 eV. In the comparison of the leakage currents in the monocrystalline and amorphous HfO_2, the leakage current in the monocrystalline HfO_2 is larger than the one in the amorphous HfO_2 for low bias, while the leakage current in the amorphous HfO_2 becomes larger as the bias increases. For low bias, the inelastic tunneling requires more phonons in the amorphous HfO_2 as compared with the monocrystalline HfO_2, because the energy differences between the traps are zero in the monocrystalline HfO_2. For high bias, the number of phonons for the inelastic tunneling process increases linearly as the electric field increases in the crystalline HfO_2, while the tunneling paths requiring fewer phonons can be found in amorphous HfO_2 where the trap levels vary over space.

In comparison of the leakage currents in the monocrystalline and amorphous HfO2, the leakage current in polycrystalline HfO_2 is larger for the bias below 1.5 V, while the averaged leakage currents are almost identical for both cases when the bias gets higher. For high bias, the single-trap assisted tunneling processes, i.e. electrode-to-trap and trap-to-electrode tunneling, dominate the leakage current. Thus, both leakage currents of amorphous and polycrystalline HfO_2 are similar. However, for low bias, in the polycrystalline HfO_2, the leakage current is dominated by trap assisted tunneling which is the trap-to-trap tunneling process because of smaller trap-to-trap distances on the grain boundaries. It results in larger leakage current in the polycrystalline HfO_2 than one in the amorphous HfO_2.

For this simplified example, the capacitor leakage is significantly lower than the transistor leakage, although this may not hold true for more realistic structures and with advanced scaling. Therefore, this KMC analysis represents an important step for the accurate optimization of the DRAM tREF by means of a TCAD-based DTCO platform.

3.3. Cell Array RC Extraction

In the previous sections we have shown how to evaluate the transistor ON-current and leakage and their stochastic dispersions. These TCAD data can be brought to SPICE level via a compact model and a circuit simulation can be performed to obtain outputs such as the DRAM writing time or refresh time. However, this task cannot be achieved without an accurate extraction of the RC parasitics, including bitline (BL) capacitance and the world line (WL) resistance. The cell array capacitance and resistance extraction are performed by using Raphael FX [22], a 3D field solver, therefore offering the highest accuracy for the RC extraction. Moreover, thanks to distributed processing (DP), the tool can keep run-time at optimal levels enabling, for example, the RC extraction of large areas within hours (instead of days). The resistance extraction accuracy is also increased by including surface scattering effect that will lead to an increased resistivity when metal lines cross-sections are scaled down.

Figure 6 shows the cell Array RC extraction flow starting from a layout-based structure generation by means of Process Explorer. Clips are

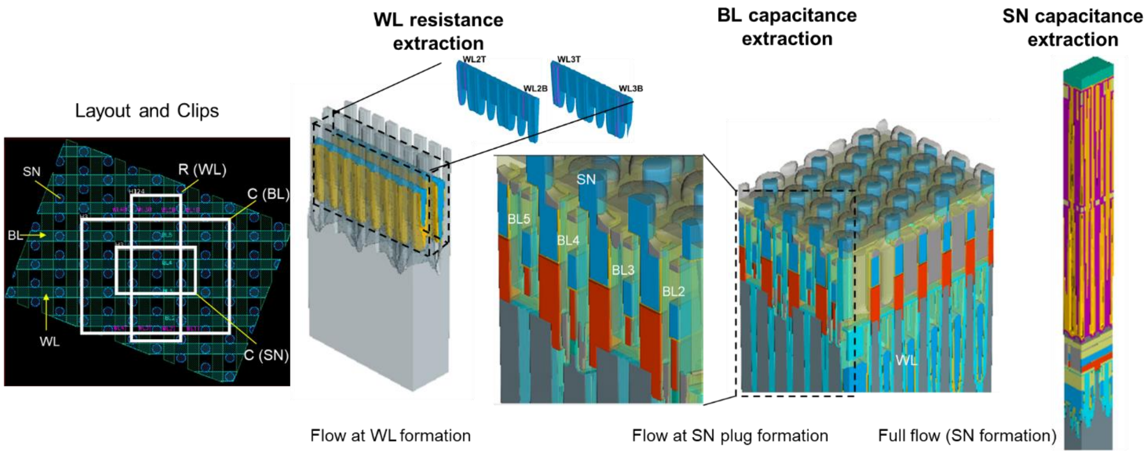


Figure 6. Cell Array RC Extraction. The extraction flow starts from a layout-based structure generation by means Process Explorer. Clips are user-specified to identify the domains of RC extraction, which is then performed by Raphael FX.

user-specified to identify the domains of the RC extraction, which is then performed by Raphael FX. Table 3 reports single cell capacitance and resistance extracted values. It is worth noting that the BL to SN capacitance dominates the total (~100aF), whilst the BL to BL coupling is relatively weak (~1aF) and the BL to WL coupling is negligible (0.01aF). The WL resistance is around 17 Ohms across the area of extraction. These results will be included in the statistical SPICE analysis presented at the end of Section 4.

Table 3. Single-cell Capacitance and Resistance extracted values.

Bit Line Capacitance Extraction		C [F]
BL3	BL2	1.54 $\times 10^{-18}$
BL3	BL4	8.48 $\times 10^{-19}$
BL3	BL5	8.95 $\times 10^{-22}$
BL3	SN	**1.23 $\times 10^{-16}$**
BL3	WL2T	2.03 $\times 10^{-20}$
BL3	WL4T	2.13 $\times 10^{-20}$
Total Capacitance		**1.26 $\times 10^{-16}$**
Word Line Resistance Extraction		R [Ω]
WL2B	WL2T	16.9
WL3B	WL3T	16.8

4. Periphery DTCO Analysis

In this section we present a TCAD-to-SPICE methodology for the early SPICE model extraction and performance evaluation of the DRAM CMOS periphery. We will focus our analysis on the Sense Amplifier (SA) circuitry, whose performance will determine the read operation reliability and, ultimately, the tREF margin.

Global variations could be modeled via different process splits accounting for the systematic variations in implant dose, geometrical dimensions and layout dependent effects – as already presented for the DRAM memory transistor in Section 3. However, because the Sense Amp performance will be mainly determined by the transistor local threshold voltage (Vth) mismatch, in the following we are going to consider only source of local statistical variability. This assumption will not distort the analysis results, unless for that cases where the process variation and local variation are highly correlated. Figure 7 shows the layout-based 3D generation of the DRAM periphery, which is achieved by means of Process Explorer [17]. S-Process [18] is employed for accurate doping and stress simulation, whilst S-Device [19] is used to generate the reference I-V and C-V characteristics that are used for the compact model extraction of the nominal device. A bulk MOSFET technology featuring a nominal gate length of 32nm and a width of 200nm is used a test-bed for this analysis.

4.1. Periphery CMOS Transistors – Statistical Variability

To account for local variability, we deploy the variability engine Garand VE [20]. In a first stage, Garand VE is calibrated against the reference I-V curves from S-Device. This includes density gradient (DG) quantum corrections, inversion charge calibration and mobility model calibration. Then all major sources of local variation are physically modelled by running hundreds statistical instances of the nominal device. These sources include random discrete doping (RDD), line edge roughness (LER) and metal gate granularity (MGG) variability (if metal gate technology) or polysilicon gate

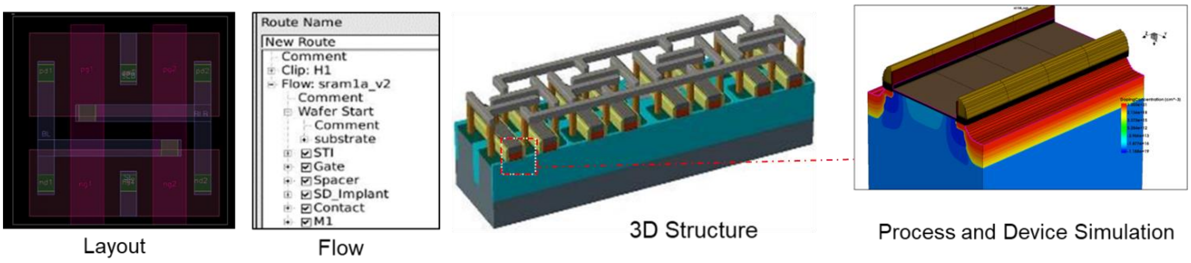


Figure 7. Layout to Process and Device simulation for the CMOS periphery Sense Amplifier. A 32nm bulk technology is considered in this example.

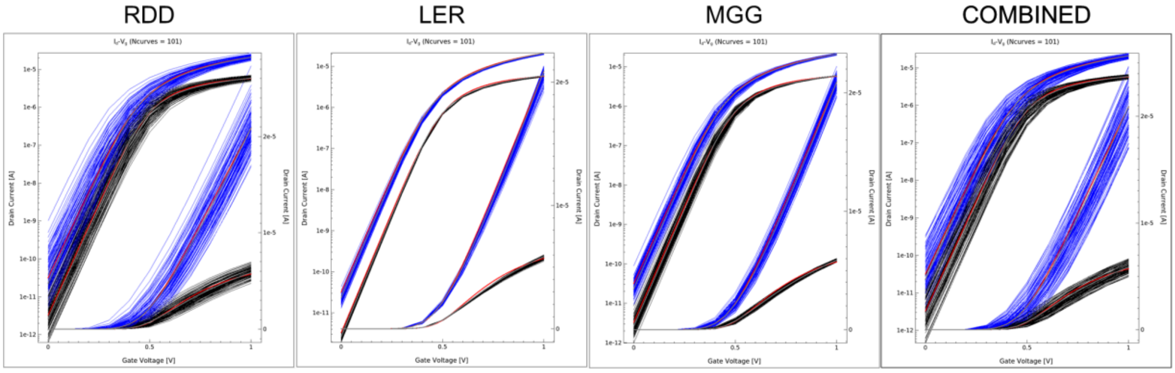


Figure 8. TCAD variability analysis considering separate and combined variability sources (RDD, LER, MGG). Results are for a width of W=25nm. The Sense Amplifier will have transistors featuring W=200 and the variability will be scaled inversely proportional to sqrt(WL).

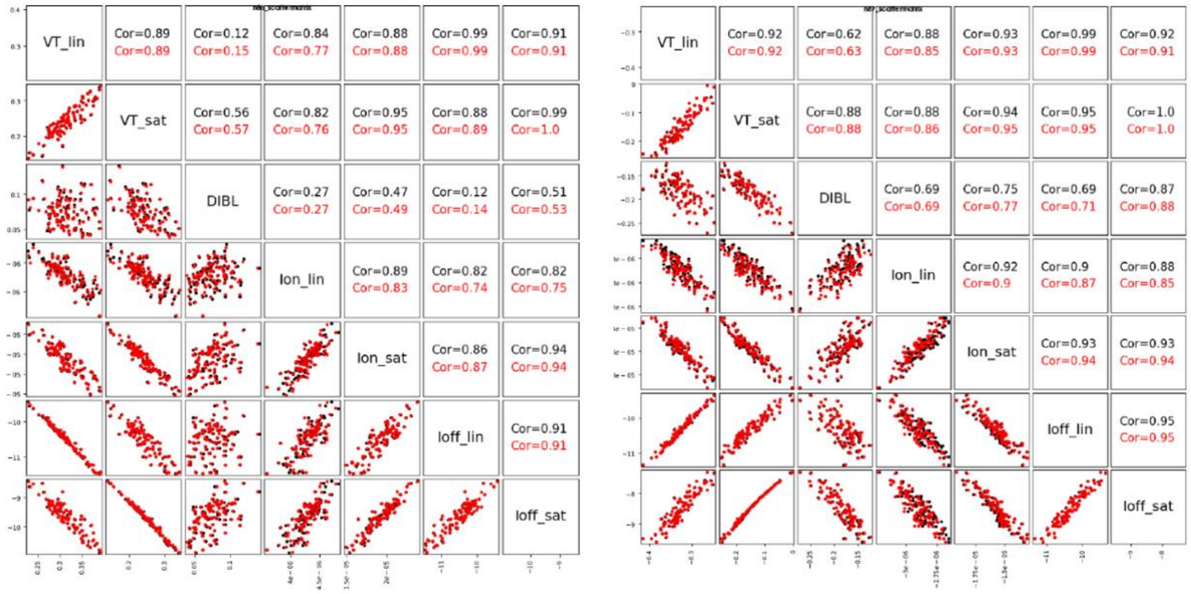


Figure 9. Compact Modelling extraction for NMOS and PMOS (RDD, LER and MGG combined). TCAD data in black and compact model results in red.

granularity (PGG) variability (if polysilicon gate technology) [33]. Figure 8 shows the I-V curves for separate and combined variability sources, highlighting that RDD and MGG play the dominant role in determining the threshold voltage and ON-current variations accounting to 15mv and 0.76μA (@W=0.2um), respectively.

Once all the target I-V/C-V characteristics are generated using physical TCAD simulation, hierarchical compact models can be extracted by

Table 4. Sense Amp Interconnect Capacitance and Resistance extracted values.

Capacitance Extraction for SPICE			C [F]
C_19_5	SEB	nmT23	6.02 ×10^{-19}
C_3_20	BLB	0mT25	6.97 ×10^{-19}
C_6_18	PG1	2mT26	5.31 ×10^{-18}
…	…	…	…
Resistance Extraction for SPICE			R [Ω]
R_0_1	ng2	0nmT18	1.50
R_2_3	IIUW2UT24	BLB	3.79
R_17_6	ng1	pg1	29.2
…	…	…	…

means of a two-stage process, involving: i) the extraction of 'uniform' or 'base' SPICE model; ii) local 'statistical' models extraction using a carefully selected subset of the compact model parameters, as detailed in [34]. The results of the extraction 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.

4.2. Periphery CMOS Interconnects – RC Extraction

Similarly to the methodology performed for the RC extraction of the DRAM cell array, Raphael FX [22] is deployed to extract the interconnect RC for the 3D structure generated by Process Explorer [17] (Figure 7). The output is a RC netlist in a SPICE-ready format which can be imported, together with the transistor models, into the statistical circuit simulator RandomSpice [23]. Table 4 shows only few lines of the extracted RC netlist.

4.3. Statistical Circuit Analysis

The simulated TCAD data is propagated into statistical SPICE models via the compact modelling extraction presented in the previous sections. The metal lines capacitive and resistive element are also added to the final netlist. For each Monte-Carlo instance of the DRAM cell, a unique leakage current is generated using the fitted TCAD data distributions. These randomized leakage values are converted to BSIM4 junction leakage parameters. The leakage compact models can reproduce the statistical TCAD data at arbitrary trap densities and storage node voltages, as verified in [28]. It is worth to remark that RandomSpice [23] directly generates the leakage values for the DRAM transistor: because we are focusing on a statistical tail analysis, the HSPICE [24] simulations can be limited to the circuits where the DRAM cell leakage current is greater than a threshold limit (here >1 fA). As a result, only ~400k out of 10M generated circuits (representing roughly 10Mbit) are run through HSPICE – enabling a very accurate, yet efficient, high-sigma analysis.

To approximate tREF through SPICE simulation, we combine the output from the SA analysis with DRAM cell analysis. The SA variability is important as it defines how much differential is required between the sensing BL and the reference BL. Local MOSFET mismatch can "offset" a SA towards one state or another, and the natural solution to this is to utilize larger devices in this circuit. However, a larger SA means that proportionally, less of the wafer area is memory cells, reducing overall memory density and increasing cost.

Utilizing the variability aware SPICE models previously extracted we can explore the tradeoff between device width and SA offset voltage as show in Figure 11 (left). In this case we select a W=200nm SA design, which leads to 48mV 3σ offset. We can then determine the minimum storage capacitor voltage required to produce a 48mV delta in the BL voltage. In this case, as shown in Figure 11 (right), 0.78V must be present on the storage capacitance in order for the '1' state to be correctly detected by a 3σ sense amp. Finally, this voltage can be plugged into the write-and-hold DRAM cell simulations.

Initial simulations, in Figure 11 (left) show the output of a 1e7 sample of cells, where process conditions are kept "nominal". Here the only variations which are applied relate to RDD and RDD+TAT interactions, and tREF at 1e-7 probability comes out at ~200ms. Finally, we also randomize process conditions for the DRAM cell – in this case this is in the form of (Dose, WLetch) variation. Each datapoint here corresponds to a 1e-7 probability cell, mixed with a 3σ sense-amp to extract a tREF distribution *per-10Mb array*. The results, in Figure 11 (right) show that, although nominal 10Mb array tREF is ~200ps, array to array tREF 1σ is ~15ms. Although the resultant tREF large compared to reported tREF values– it is worthwhile noting that this analysis was performed at 27C, and

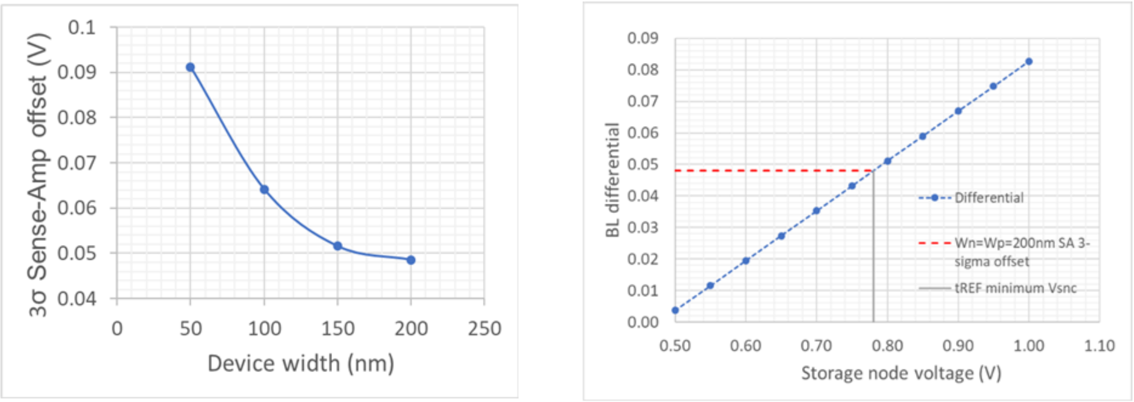


Figure 10. (left) Sense-amp offset analysis, showing offset vs. nMOS/pMOS device size. (right) Determination of minimum storage node voltage required to correctly sense the '1' state of the capacitor.

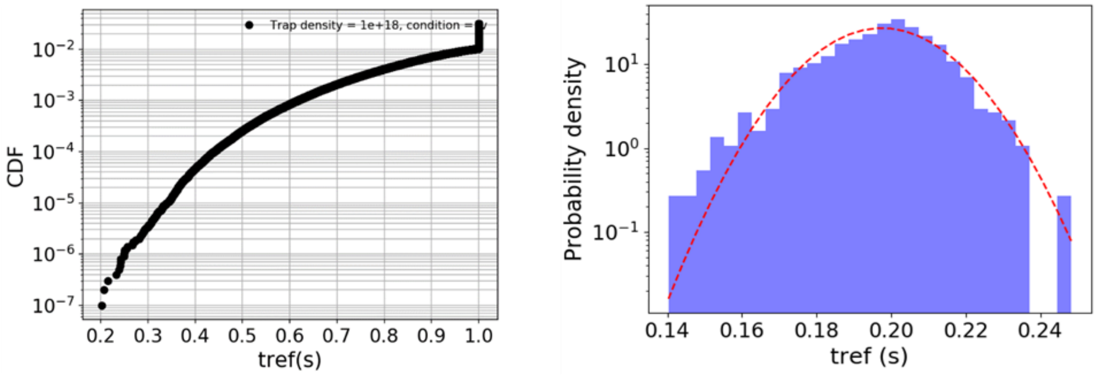


Figure 11. (left) tREF tail at a nominal process condition, showing how long it takes for Vsnc to drop to 0.78V. (right) Distribution of tREF produced at 1,000 different random process conditions – effectively measuring tREF from 1,000 ~10Mb arrays.

not worst-case temperature, where tREF time can easily drop by a significant factor up to 0.3, when shifting from 27C to 80C [35]. Final, these results can be compared to tREF/yield specifications for the process – if yield targets are not achieved, updates in the design may be considered. For example, resizing or redesigning of the sense-amp, to reduce the BL differential requirements and increase tREF can be quantitatively evaluated. This, and other process updates can be quickly evaluated by rerunning the flow with updated inputs.

5. Conclusions

The semiconductor industry is facing a paradigm shift, with scaling being now driven by more frequent technology releases for both memory and logic. DTCO methodology becomes the key to unlock the potential of each release, by means of the efficient and accurate exploration of different technological variations and the optimization of fundamental figures of merit such as Power-Performance-Area-Cost, memory cell retention time, and parametric yields. In this paper we have presented a DTCO analysis of an advanced DRAM technology, aiming at the optimization of the DRAM refresh time. In particular, we have shown how the several components affecting the memory and the logic part can be captured by a multi-stage simulation approach including both process and statistical variations. This enables a variability-aware DTCO particularly suited for optimizing performance and yields of advanced memory technologies, reducing manufacturing cost and cycle time and accelerating time-to-market.

References

[1] V. Moroz, X.-W. Lin, T. Dam, "Logic Block Level Design-Technology Co-Optimization is the New Moore's Law", 2020 4th IEEE Electron Devices Technology & Manufacturing Conference (EDTM).

[2] P. Matagne, H. Nakamura, M.-S. Kim et al., “DTCO and TCAD for a 12 Layer-EUV Ultra-Scaled Surrounding Gate Transistor 6T-SRAM”, 2018 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD).

[3] A. Asenov, B. Cheng, X. Wang et al., “Variability Aware Simulation Based Design- Technology Cooptimization (DTCO) Flow in 14 nm FinFET/SRAM Cooptimization”, IEEE Transaction on Electron Devices, pp.1682-1690, vol.62, 2015.

[4] Z. Zhang, R. Wang, C. Chen et al., “New-Generation Design-Technology Co-Optimization (DTCO): Machine-Learning Assisted Modeling Framework”, 2019 Silicon Nanoelectronics Workshop (SNW).

[5] A. Asenov, K. El Sayed, R. Borges et al., “TCAD based Design-Technology Co-Optimisations in advanced technology nodes”, 2017 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA).

[6] S.C. Song, B. Colombeau, M. Bauer et al., "2nm Node: Benchmarking FinFET vs Nano-Slab Transistor Architectures for Artificial Intelligence and Next Gen Smart Mobile Devices", Symposium on VLSI Technology, pp. 206-207, 2019.

[7] J. X. Niu, H. Veluri, A. V.-Y. Thean, “Design-Technology Co-optimization (DTCO) for Emerging Disruptive Logic & Embedded Memory Process Technologies”, 2019 Electron Devices Technology and Manufacturing Conference (EDTM).

[8] Y. Kim, U. Monga, J. Lee et al., “The efficient DTCO Compact Modeling Solutions to Improve MHC and Reduce TAT”, 2018 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD).

[9] I. Jang, H. Ko, A. Schmidt et al., “Multi-domain process modeling for advanced logic and memory devices: from equipments to materials”, 2018 IEEE International Electron Devices Meeting (IEDM).

[10] S.-H. Lee, “Technology scaling challenges and opportunities of memory devices”, 2016 IEEE International Electron Devices Meeting (IEDM).

[11] S.-W. Park, S.-J. Hong, J.-W. Kim et al., “Highly Scalable Saddle-Fin Transistor for Sub-50nm DRAM Technology”, 2006 Symposium on VLSI Technology, 2006. Digest of Technical Papers.

[12] S.-W. Ryu, K. Min, J. Shin et al; “Overcoming the reliability limitation in the ultimately scaled DRAM using silicon migration technique by hydrogen annealing”, 2017 IEEE International Electron Devices Meeting (IEDM).

[13] C.-M. Yang, C.-K. Wei, Y. J. Chang et al; “Suppression of Row Hammer Effect by Doping Profile Modification in Saddle-Fin Array Devices for Sub-30-nm DRAM Technology”, IEEE Transactions on Device and Materials Reliability, pp. 685-687, vol.16, 2016.

[14] S. H. Jang, J. Lim, J. Han et al., “A Fully Integrated Low Voltage DRAM with Thermally Stable Gate-first High-k Metal Gate Process”, 2019 IEEE International Electron Devices Meeting (IEDM).

[15] S-K Park, “Technology Scaling Challenge and Future Prospects of DRAM and NAND Flash Memory”, 2015 IEEE International Memory Workshop (IMW).

[16] M. H. Cho, N. Jeon, T. Y. Kim et al., “An Innovative Indicator to Evaluate DRAM Cell Transistor Leakage Current Distribution”, IEEE Journal of the Electron Devices Society, pp. 494-499, vol. 6, 2018.

[17] *Process Explorer User's Manual v. R-2020.09 Synopsys*, 2020.

[18] *Sentaurus Process User's Manual v. R-2020.09 Synopsys*, 2020.

[19] *Sentaurus Device User's Manual v. R-2020.09 Synopsys*, 2020.

[20] *Garand VE User's Manual v. R-2020.09 Synopsys*, 2020.

[21] *Mystic User's Manual v. R-2020.09 Synopsys*, 2020.

[22] *Raphael FX User's Manual v. R-2020.09 Synopsys*, 2020.

[23] *RandomSpice User's Manual v. R-2020.09 Synopsys*, 2020.

[24] *HSPICE User's Manual v. R-2020.09 Synopsys*, 2020.

[25] A. Ghetti, C. Monzio Compagnoni, L. Digiacomo et al., “Evidence for an atomistic-doping induced variability of the band-to-band leakage current of nanoscale device junctions”, 2012 International Electron Devices Meeting.

[26] T. Yang, X.-W. Lin, “Trap-Assisted DRAM Row Hammer Effect”, IEEE Electron Device Letters, pp. 391-394, vol. 40, 2019.

[27] I. Ciofi, P. J. Roussel, C. J. Wilson et al., “Variability-Aware Predictive Modeling of Line-to-Line Dielectric Reliability”, IEEE Transactions on Electron Devices, pp. 1737-1744, vol. 67, 2020

[28] S. M. Amoroso, J. Lee, P. Asenov et al., “High-sigma analysis of DRAM write and retention performance: a TCAD-to-SPICE approach”, 2020 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD).

[29] G.C. Jegert, “Modeling of Leakage Currents in High-k Dielectrics,” Ph.D. Thesis, Technishen Universität München, Sept 9, 201l.

[30] M. Herrmann and A. Schenk, “Field and High-temperature Dependence of the Long Term Charge Loss in Erasable Programmable Read Only Memories: Measurements and Modeling,” J. Appl.Phys., vol. 77, no. 9, pp. 4522-4540, May 1995.

[31] J. Frenkel, “On pre-breakdown phenomena in insulators and electronic semiconductors”, Physical Review., vol. 54, no. 8, pp. 647–648, 1938.

[32] L. Vandelli, A. Padovani, L Larcher et al., “Microscopic Modeling of Electrical Stress-Induced Breakdown in Poly-Crystalline Hafnium Oxide Dielectrics”, IEEE Transactions on Electron Devices, vol. 60, pp. 1754-1762, 2013.

[33] A. Cathignol, B. Cheng, D. Chanemougame et al., “Quantitative Evaluation of Statistical Variability Sources in a 45-nm Technological Node LP N-MOSFET”, IEEE Electron Device Letters, vol. 29, 2008.

[34] X. Wang, B. Cheng, D. Reid et al., "FinFET Centric Variability-Aware Compact Model Extraction and Generation Technology Supporting DTCO", IEEE Transactions on Electron Devices, vol. 62, pp. 3139-3146, 2015.

[35] J. Liu, B. Jaiyen, Y. Kim, C. Wilkerson, O. Mutlu, “An experimental study of data retention behavior in modern DRAM devices: implications for retention time profiling mechanisms” ISCA 2013, Proc. of the 40th Annual International Symposium on Computer Architecture, pp. 60-71, 2013

Photography & Biography

Salvatore Maria Amoroso received his Ph.D. in Electronic Engineering from Politecnico di Milano in 2012. He has been with the Device Modelling Group of University of Glasgow as a Research Associate, working on the advanced simulation of variability and reliability of decananometer MOSFETs and Flash Memories, until 2014. He joined Gold Standard Simulations Ltd, in 2014 working as a Senior Engineer for TCAD software development and customer accounts manager. Since 2016 he is with Synopsys, Inc. working as an R&D Engineer on the development of advanced TCAD-to-SPICE methodologies and Design-Technology Co-Optimization (DTCO) enablement.

Plamen Asenov received is PhD from Glasgow University in 2012. He has been with ARM, where he worked on embedded memory design at advanced nodes until 2015. He joined Gold Standard Simulations in 2015, working on DTCO applications. Since 2016 he has been with Synopsys as an R&D Engineer, specializing in TCAD-to-SPICE and DTCO across a range of technologies.

Jaehyun Lee has been with SK Hynix from 2008 to 2012 as Research Engineer for mobile DRAM development. He received his Ph. D. in Electrical Engineering from Korea Advanced Institute of Science and Technology in 2016. He joined Device Modelling Group of University of Glasgow, working as a Research Associate on the software development of interconnect and nanoscale MOSFETs. Since 2018, he is with Synopsys working as an R&D Engineer on the development of TCAD-to-SPICE and Design-Technology Co-Optimization (DTCO) methodologies.

Nara Kim has received the B.S degree in physics from Konkuk University, Seoul, South Korea, in 2009. She has been with Semiconductor Research and Development Center, Samsung Electronics Company Ltd., Hwaseong, South Korea as TCAD engineer to research on the development of the semiconductor device from 2009 to 2019. Since 2019 she has been with Synopsys, Inc. working as an Application Engineer.

Yong-Seog Oh joined Daewoo Telecom in Korea to develop BiCMOS technology as soon as he received B.S. in physics from Seoul National University in 1987. He led the team for TCAD simulation and SPICE parameter extraction as well as the test pattern design including ESD protection. In 1994, he moved to US to join Stanford spin-off Technology Modeling Associates (TMA) as an R&D engineer for the process simulator, TSUPREM-4. After IPO in 1997, TMA became a part of Avant! in 1998 and of Synopsys in 2002. He accomplished many projects on process and material model development until 2019. His current main concern is on the device reliability, topography modeling and calibration for the state-of-art semiconductor process.

Lee Smith received the B.S. degree in physics from the University of Florida and the Ph.D. degree in physics from Stanford University. Dr. Smith is currently R&D manager of the TCAD Device Simulation group at Synopsys which is engaged in the development of advanced device modeling techniques for a variety of applications including CMOS, power, RF, and memory devices.

Xi-Wei Lin is Business Development Director from Silicon Engineering Group in Synopsys, currently focusing on TCAD and DTCO applications for logic and memory technology developments. He previously worked at Micron Technology, LSI Logic, Philips Semiconductors, and Lawrence Berkeley National Laboratory, responsible for materials science and engineering, CMOS process technology development, ASIC and memory designs and verification, as well as power methodology and standard cell library architecture. He received his B.S. degree in microelectronics from Beijing University, China and M.S. and Ph.D. in solid state physics from University of Paris, Orsay, France.

Victor Moroz received M.S. degree in Electrical Engineering from Novosibirsk Technical University and Ph.D. degree in Applied Physics from the University of Nizhny Novgorod. After engaging in technology development at several semiconductor manufacturing companies and teaching semiconductor physics at a University, Dr. Moroz joined a Stanford spin-off Technology Modeling Associates in 1995. After IPO in 1997, the TMA TCAD team became part of Avanti in 1998, and in 2002 it became a key part of Synopsys, connecting a synthesis company to the manufacturing. Currently Dr. Moroz is a Synopsys Fellow, engaged in a variety of projects on modeling advanced CMOS with over 100 US patents, and serving as an Editor of IEEE Electron Device Letters.

J. Microelectron. Manuf. **3**, 20030410 (2020)
doi: 10.33079/jomm.20030410

Pattern-Centric Computational System for Logic and Memory Manufacturing and Process Technology Development

Chenmin Hu, Khurram Zafar, Abhishek Vikram*, Geoffrey Ying

Anchor Semiconductor, Santa Clara, USA, 95051

Abstract: Chip designers employ computer-aided design, circuit simulation, and design rule check systems. Lithography engineers employ model-based OPC (Optical Proximity Correction) and model-based print-simulation systems. Reticle inspection teams employ Aerial Image Measurement Systems® and Virtual Stepper® Systems. These teams are accustomed to evaluating and deploying state-of-the-art computational systems. When real-silicon fabrication begins, however, the teams responsible for line monitoring, wafer inspection, and yield attainment operate without the benefit of similarly advanced computational systems. In this paper we describe such a system and explore its applications and benefits. The *system* has received three U.S. patents [1-3] and brings together the significant potential of CAD (Computer Aided Design) layout (GDS, OASIS), Die-to-Database, and Machine Learning to build a dynamic, self-improving computational system. Featuring care area generation, advanced machine learning-based SEM (Scanning Electron Microscope) sampling that optimizes both DOI (Defect of Interest) capture rate and discovery of new defect types, comprehensive extraction of all *Information of Interest (IOI)* from all SEM images, detection of defect types not possible before, massive pattern fidelity analysis, full chip pattern decomposition and risk scoring via machine learning, innovative PWQ (Process Window Qualification) analysis and process window determination, risk assessment of new tape-outs, large scale in-wafer OPC verification and more, the *system* delivers a comprehensive *pattern centric* platform for process technology development and manufacturing.

Keywords: Die-to-Database, Full Chip Decomposition, Machine Learning, Defect Discovery, Pattern Fidelity, Pattern Risk Scoring, OPC Verification, Process Window Qualification.

1. Introduction

At every major technology node, the density of transistors per unit area approximately doubles, and so does the quantity of *raw data* that fabs need to extract, track, and analyze. Compounding the problem is the fact that doubling the density of transistors means shrinking their size. Not only are smaller geometries harder to fabricate, they are harder to inspect. The semiconductor industry has witnessed a rapid progression of technology nodes thanks to advancements in lithography such as ArF Immersion and EUV (Extreme Ultra Violet wavelength), and attendant advancements in material stacks. These advancements have precipitated advancements in adjacent areas. For the areas of wafer inspection, line monitoring and yield enhancement, adjacent advancements have been made in E-Beam (electron beam) and SEM technologies that have the ability to detect and resolve increasingly smaller deviations in increasingly smaller geometries – and at relatively higher speeds. However, these tools are still throughput-limited and fabs continue to employ a combination of (a) high-speed *low*-resolution optical tools and (b) low-speed *high*-resolution E-Beam and SEM tools.

Driven by market and technology demands, leading manufacturers of E-Beam and SEM tools are investing aggressively in new technologies such as faster single-beam systems (that feature larger spot sizes while retaining high resolutions) and multi-beam systems to confront the continuing challenges of throughput and coverage. But hardware alone is not sufficient for yield learning and line monitoring. Hardware generates raw data, but not *information*. Software generates *information* and, more importantly, *actionable information*.

In this paper we present a pattern-centric computational system for the fab that leverages the fields of CAD layout (GDS/OASIS), Die-to-Database, and Machine Learning to enable bold new opportunities for wafer inspection, SEM review, defect discovery, (Focus Exposure Matrix) FEM/PWQ analysis, Litho/OPC optimization, pattern fidelity monitoring, yield prediction and risk

[*] Address all correspondence to Abhishek Vikram, E-mail: abhishek@anchorsemi.com

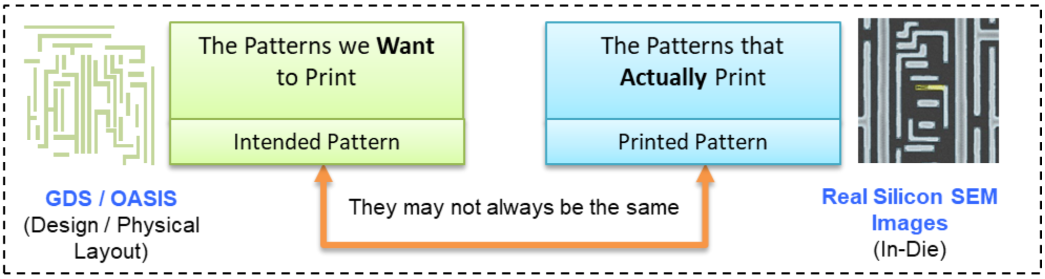

The Patterns we **Want** to Print
Intended Pattern
The Patterns that **Actually** Print
Printed Pattern
GDS / OASIS
(Design / Physical Layout)
They may not always be the same
Real Silicon SEM Images
(In-Die)

Figure 1. Intended Patterns are represented in the CAD Layout, and Printed Patterns are represented in SEM images, from which the Printed Pattern Database is built.

assessment (especially for new tape-outs), and more. We begin with a brief discussion of the technology before focusing on value-added applications.

2. Building a Pattern Centric Computational System

Because the essential task of a semiconductor wafer fab is to print *patterns* onto the wafer, Anchor's computational system is designed to be *pattern centric*. The CAD layout is a database of patterns. OPC is performed on patterns. Mask writers etch patterns (contained in MEBES files). Lithography process windows are determined using FEM/PWQ techniques that analyze patterns in each focus/exposure modulation. Test chips are composed of a diversity of patterns. DFM (Design for Manufacturing) databases record weak patterns. DRC (Design Rule Check) rule decks are designed to avoid problematic pattern layouts.

Patterns are indeed essential components. But the notion of patterns takes a back seat in the operation of the wafer fab. This is not necessarily *desirable*, but it is *understandable* because (a) the design house is CAD based, (b) the OPC team is CAD based, and (c) the mask house is CAD based. But not the fab. Once the reticle or mask enters the fab, the *digital* side of manufacturing is essentially complete (where every digital "run" produces *identical* results), and the *analog* side begins (where every analog "run" produces slightly *different* results). Like snowflakes, no two wafers nor any two die are exactly alike. There are differences every time the wafer is exposed or developed or etched or planarized or implanted or cleaned. The process steps leading from the front end of line to the back end of line are *analog* steps.

For a fab operating in the *analog* domain to communicate and coordinate more effectively with the Design, OPC and Mask teams that operate in the *digital* domain, it needs to adopt the language of patterns as well.

For years, fabs have struggled to cope with patterns, often spending days or weeks of manual effort to analyze large quantities of FEM/PWQ results, for example, and provide actionable information to the OPC team or to appropriate process modules.

Anchor's computational system arises from the intersection of the two primary domains of *intended* and *printed* patterns shown in *Figure 1*, and consists of three pillars:

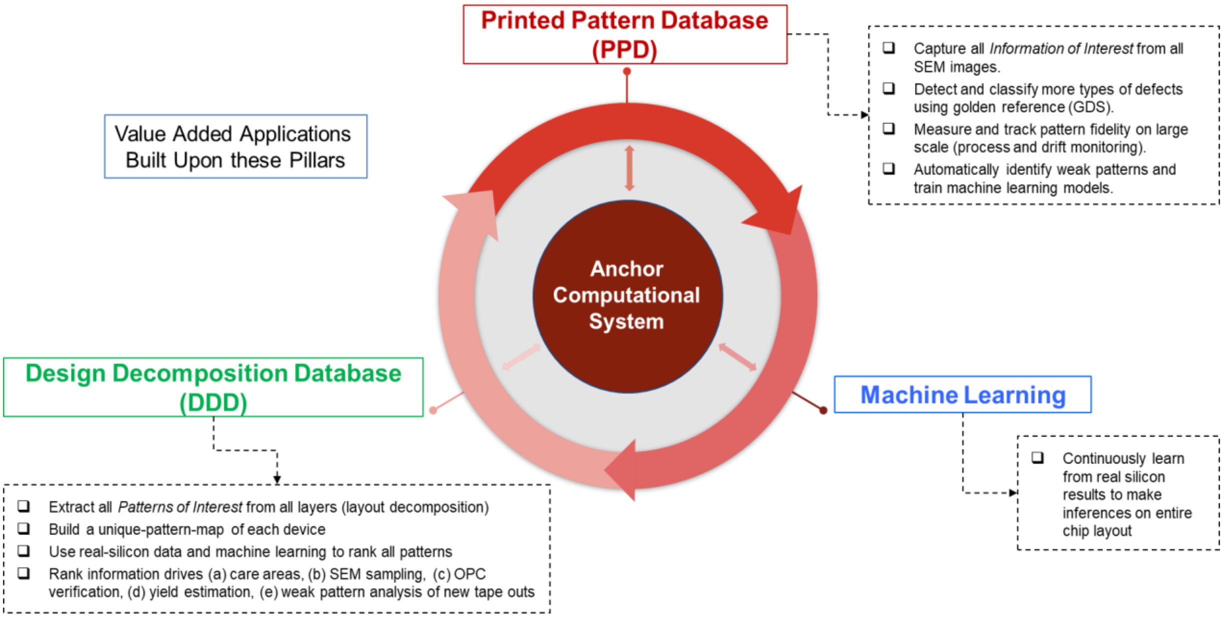
1. Printed Pattern Database
2. Design Decomposition Database
3. Machine Learning

3. Three Pillars of a Computational System for the Fab

The CAD layout is a *golden reference* database of the intended patterns. Over the past decade and a half, use of CAD inside the fab has enabled new opportunities for yield analysis and wafer inspection. But is there an *analog* equivalent of the CAD layout? That is, is there a database of the *printed* patterns?

As shown in *Figure 2*, if a database of printed patterns were to exist, it could once again enable new opportunities for process technology development and manufacturing. We call this the *Printed Pattern Database*, and it is the most fundamental of the three pillars of Anchor's computational system.

The printed pattern database is constructed in an intelligent manner that extracts and retains only the patterns of interest within each SEM image. Patterns of interest are identified by a set of *parametric search rules* that operate in real time on each image. Once extracted, each pattern of interest is assigned a class code corresponding to the *rule* that identified the pattern. For example, when a tip-to-line pattern is found, it is classified as a tip-to-line. When a tip-to-tip pattern is found, it is classified as a tip-to-tip. This enables the user to query and study the yield


Printed Pattern Database (PPD)
Value Added Applications Built Upon these Pillars
Anchor Computational System
Design Decomposition Database (DDD)
Machine Learning

□ Capture all *Information of Interest* from all SEM images.
□ Detect and classify more types of defects using golden reference (GDS).
□ Measure and track pattern fidelity on large scale (process and drift monitoring).
□ Automatically identify weak patterns and train machine learning models.

□ Extract all *Patterns of Interest* from all layers (layout decomposition)
□ Build a unique-pattern-map of each device
□ Use real-silicon data and machine learning to rank all patterns
□ Rank information drives (a) care areas, (b) SEM sampling, (c) OPC verification, (d) yield estimation, (e) weak pattern analysis of new tape outs

□ Continuously learn from real silicon results to make inferences on entire chip layout

Figure 2. Three Pillars of Anchor's Computational System.

impacts of specific types of patterns along with the *variations* of those patterns (e.g. study the differences in printability of tip-to-line patterns as a function of the gap between tip and line).

The *Design Decomposition Database* is the next pillar. Each layer of interest in the CAD Layout is decomposed systematically into a collection of unique constituent patterns of a specified maximum size. A poly layer, for instance, will be fully decomposed into its unique constituent patterns.

Decomposition is performed using the same parametric rule engine that builds the printed pattern database, which means that only *patterns of interest* are extracted when decomposing the full chip layout. This eliminates *don't care* patterns that would otherwise burden the database with too many nuisance patterns. When a layer is fully decomposed into its constituent patterns of interest, the result is an abbreviated representation of the layer.

The third pillar, *machine learning*, bridges the first two pillars and enables entirely new opportunities for yield learning and process optimization.

There are at least two ways to model a real-world system in order to make specific kinds of predictions. The conventional method is to build the model from first principles and tune the model until it begins to make sufficiently accurate predictions. This is done, for example, with OPC Simulation where selected properties of light waves, optics, and materials are combined into a mathematical model that takes a CAD layout (post-OPC layout) as input and generates a simulated print (contours) as output [4-6]. Unfortunately, such models have considerably expensive development, optimization, and run-time requirements.

The alternative method is to allow a computer to build the model itself using appropriate training data that provide sufficient examples of *if this goes in, then that comes out*. The computer examines all of the inputs and their expected outputs and builds a self-learning model that can take a new input not seen before and *infer* or predict the output. Anchor's computational system applies this idea in many ways, one of which is to learn from the Printed Pattern Database and assess the printability risk of all patterns in the Design Decomposition Database.

The Printed Pattern Database (PPD) provides exactly the training set necessary for Machine Learning because it contains both the (a) intended pattern (in CAD database) and the (b) printed pattern (on die). This provides a rich training set because it contains numerous examples of *if this goes in (the intended pattern), then that comes out (the printed pattern)*. New SEM images that are continuously being captured by the fab are added to the PPD. This dynamic environment allows the machine learning system to learn continuously and therefore improve its prediction accuracy. As the accuracy of the machine improves over time, the system moves closer to an *expert* system.

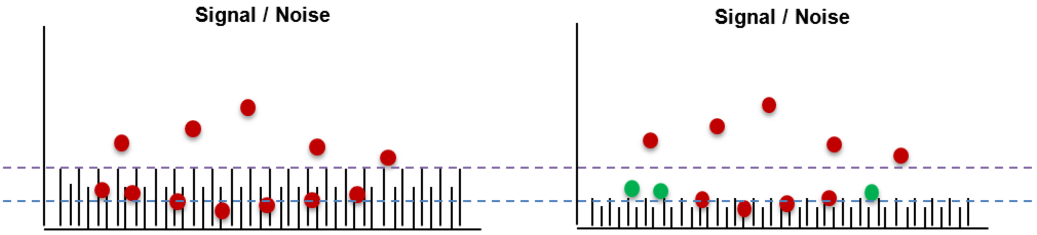


Figure 3. Noise characteristics of conventional care areas on the left may contain undetectable defects because they are buried below the noise floor. More homogeneous care areas result in less noise (right) within each care area group, allowing previously hidden defects (green) to rise above the noise floor and be detected.

4. Value-Added Applications

Numerous value-added applications are enabled by the three pillars of Anchor's pattern centric computational system. We discuss a handful of those applications at an introductory level in order to keep this paper reasonable in size.

4.1. Care Area Generation for Optical and E-Beam Inspection

Optical inspection tools are still essential because of their high throughput and high wafer coverage. Although they lack the resolving power of an E-Beam tool, they incorporate advanced features such as KLA's *NanoPoint*® / *PinPoint*® and Applied Materials' *Marker*® that attempt to improve sensitivity to defects by reducing a particular type of system noise [7, 8]. To make use of these features, it is first necessary to perform full-layer pattern segmentation such that the patterns in each segment are relatively homogeneous. Inspection recipes can be optimized for each segment, thereby improving sensitivity in each segment, shown in *Figure 3*.

Anchor's Design Decomposition Database (DDD), with its ranked collection of patterns, enables this segmentation in a manner not possible before. High-risk patterns from the DDD are first placed into "look-alike" groups such that the patterns within each group are relatively homogenous. Then each look-alike group is exploded, which means that all *instance locations* of all member patterns are added to the group. Now each group contains a set of look-alike patterns and every location on the die where those patterns occur. Each group becomes a "segment" for a KLA or Applied Materials inspection tool. These segments, consisting of relatively high-risk patterns, can be inspected with high sensitivity without incurring high noise.

E-Beam inspection tools are playing an increasing central role because of their ability to resolve tiny details on leading edge technology nodes. Although they lack speed and provide limited wafer coverage, advancements are being made to both speed and resolution. For any low-throughput tool, choosing the right care areas is of paramount importance. High risk patterns in the Design Decomposition Database are used to supplement a fab's existing E-Beam care area.

4.2. Review SEM Sample Plan

Review SEMs have been used for decades to compensate for a lack of resolution on optical inspection tools. The *patch images* they produce are pixelated and cannot be used to adequately scrutinize the properties of every defect. A clear and detailed image of the defect is necessary to determine its type, its shape, its causal mechanism, and its impact to yield (killer versus non-killer).

Because of the relatively slow throughput of Review SEM tools, it is necessary to pick a subset of the defects that were detected by the optical inspector. If a poor subset is picked or *sampled*, not much is learned. Fabs generally expect the sampled subset to (a) contain as many known defects of interest (DOI) as possible and (b) discover new defect types. It may seem straightforward to generate a sample plan that addresses both needs, but these are often competing requirements. If the sample plan is biased too much around (a), it will lose its ability to discover new defect types (b), and vice-versa.

Given a sampling budget of N defects for SEM review, Anchor's computational system generates a balanced sample plan while providing users the ability to bias the algorithm a little in either direction. Balancing the sample plan means choosing defects from the original population whose *extended* properties are likely to both (a) increase capture rate of known DOI and (b) discover new defect types. In broad terms, Anchor's computational system derives these *extended* properties and creates a final sample plan through a combination of supplied defect

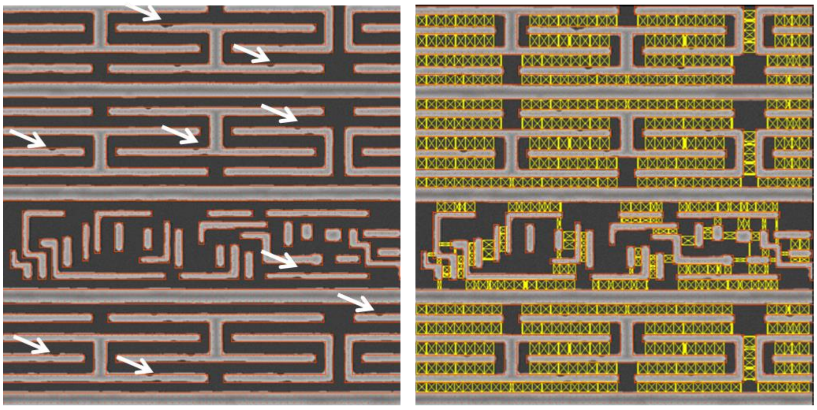


Figure 4. All Information of Interest Extracted from SEM Image. The image is scanned for both defects (left) and patterns-of-interest (right). All defects are classified and reported; and all patterns-of-interest are measured and tracked in the Printed Pattern Database.

properties, generation of new properties through reprocessing of patch images, and machine learning.

The computational system offers an additional option to apply pattern risk scores that are stored in the Design Decomposition Database. Sample plan candidates can be further filtered in or out based on their pattern risk scores.

4.3. Comprehensive Extraction of *Information of Interest (IOI)* from SEM images

Despite the paramount importance of high-resolution SEM images at all technology nodes, and especially the leading technology nodes, they are predominantly wasted. At sub-14nm nodes in particular, each SEM image contains large amounts of information, but conventional workflows examine only the center of each image to classify a defect that is expected to be present in the center. Unfortunately, upwards of 50% to 70% of SEM images do not contain a "SEM visible" defect in the center. It is likely that some sort of anomaly is indeed present in the center because the optical column in the wafer inspection tool registered an anomaly. But a SEM tool is not an optical tool; the mechanics of electron beam emission and scatter are sufficiently different from the mechanics of photon emission, transmission, and reflection. So, a SEM tool is physically unable to *see* certain types of *optical* defects, and these are referred to as SEM Non-Visuals or SNVs.

When we consider the low throughput of a SEM tool, the limited number of images that the fab's cycle time allows, and the paramount importance of the SEM for yield learning, it is profoundly disconcerting to realize that 50% to 70% of SEM images are simply discarded for being SNV and the remaining ones are examined in a superficial manner (i.e., the center portion of the image is examined for the presence of a defect, and the defect is classified). The type of information that is most effective for yield learning also happens to be the information that is most often discarded.

Anchor's Printed Pattern Database and value-added applications eliminate that waste.

Every SEM image, regardless of SNV status, is analyzed from head to toe. As shown in *Figure 4*, every bit of *Information of Interest (IOI)* is extracted and recorded in the Printed Pattern Database for the value-added applications to exploit. Parametric pattern search rules are invoked on each image to find and extract Information of Interest while rejecting *don't care* features. Information of Interest includes *named* patterns of interest and their measurements. For example, a named pattern might be a *tip-to-line* or a *comb* or a *line end with single via* or a *set of dense thin lines*, etc. Their measurements will indicate how well or how poorly each named pattern is printing – in effect, this enables *pattern fidelity monitoring and analysis*.

Each SEM image is also checked for the presence of any number of predefined defect types such as hard breaks and bridges, soft breaks and bridges, line end pullback with exposed vias, misshapen contacts and vias, etc. Conventional workflows look for one defect per image (1-to-1), but Anchor's computational system looks for all defects on each image (1-to-many). As we discuss later, Anchor's die-to-database approach for defect detection enables new types of defects to be detected.

4.4. Detection / Discovery of Defect Types not Possible or Practical Before

Conventional defect detection methods rely on target-die to reference-die comparison where the reference die may be adjacent to the target die or it

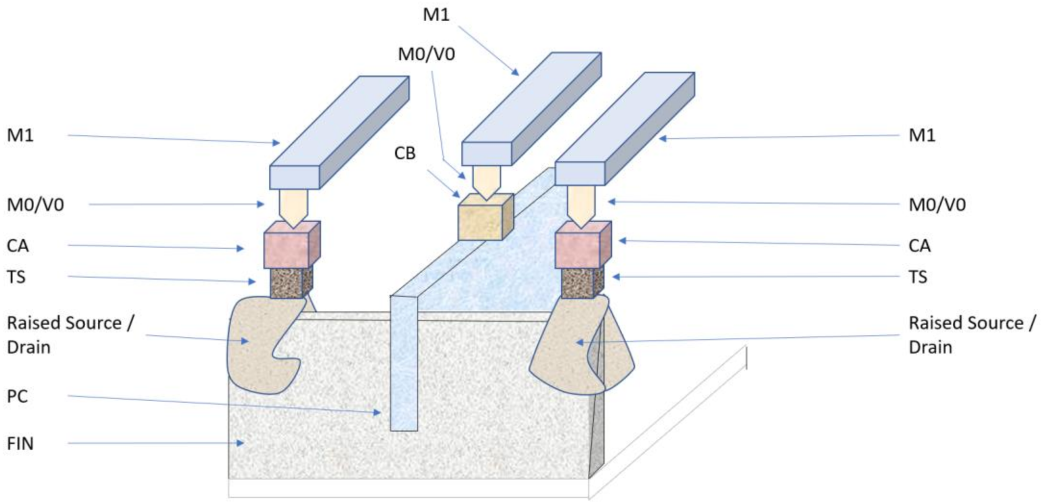


Figure 5. Representative cross-section of FinFET device.

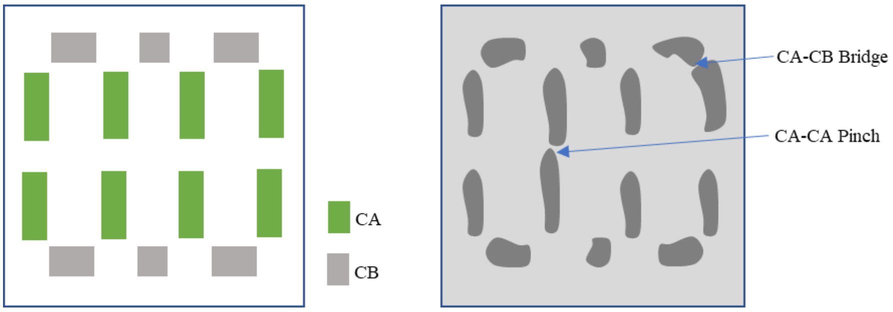


Figure 6. Design layout for Contact layers CA/CB (left) and Representative SEM image for Contact layer features (right).

may be a preselected golden die. There are several limitations with this approach that prevent certain categories of defects from being detected and corrected, leading to diminished yields and extended process debug cycles. Here we list some of the defect types that are either impractical or impossible to detect using conventional methods, but which are fully detected by Anchor's pattern centric computational system. Some of the examples in the ensuing subsections will refer to the representative cross-section shown in *Figure 5*.

4.4.1. CA to CB Bridge (Short)

CA and CB structures are part of the same *contact* layer, but they connect to different functional elements of the transistor. CA connects to *source* and *drain*, but CB connects to *poly (PC)*. Variations in the patterning process for contact layers may result in undesirable bridging between CA and CB structures. This bridging could be the result of marginalities in (a) design, (b) lithography, or (c) etch. Without the chip design serving as the reference, it is impractical for yield engineering to distinguish between CA and CB in images where only the contact layer is visible. Cross sectional analysis may be needed to positively distinguish CA from CB because such an analysis reveals the under or previous layer to which each contact is connected. Anchor's pattern-centric approach, however, can readily detect CA-CB bridges and distinguish them from CA-CA and CB-CB bridges, as shown in *Figure 6*.

4.4.2. Line End Pullback Leading to Exposed VIA/Potential VIA Disconnection (Open)

The manner in which a wafer inspection recipe is tuned or optimized can result in either a significant under-detection or over-detection of line-end pullbacks. Detection of pullbacks is essential, but not all pullbacks are killer defects or otherwise consequential. Pullbacks that *are* consequential cannot be differentiated from the entire set of pullbacks because conventional defect detection methodologies lack a comprehensive reference image from which such determinations can be made. D2DB-PM, however, uses the comprehensive chip

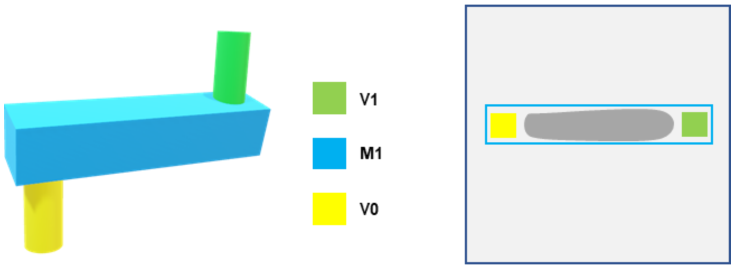


Figure 7. Target design for V0-M1-V1 overlay (left) and Representative M1 contour with design overlay. Exaggerated line-end pullback on both ends for discussion purposes (right).

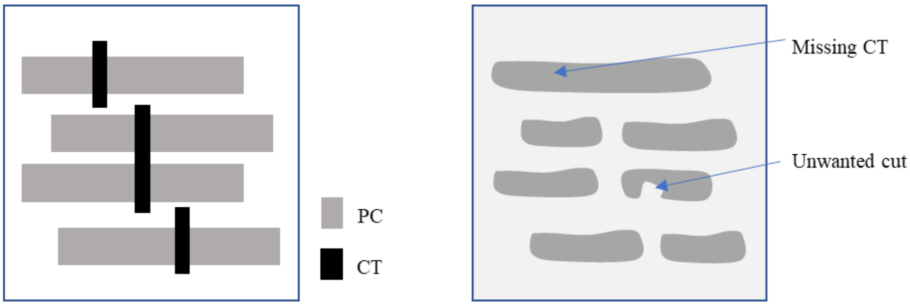


Figure 8. Target PC feature with Cut Mask (left) and Representative SEM image after cut or CT (right).

design as its reference, and is therefore able to detect additional categories of defects such as 'line-end pullback with exposed under layer via' and 'line-end pullback with *future* exposed *upper* layer via' that will result in an electrical disconnect or increased resistivity, as shown in *Figure 7*.

4.4.3. Cut Layer Issues (Short, Open)

Cut Masks are commonly used in advanced nodes to assist in printing of short lines with narrow gaps, as explained in *Figure 8*. This widely adopted method prints long lines and then *cuts* them into the desired lengths with a subsequent cut mask. But the placement or overlay of the cut mask atop the previous layer is not always optimized and may render unwanted artifacts and errors on the wafer. Without access to a comprehensive reference image, conventional defect detection methodologies are unable to (a) detect all such defects and (b) to do so reliably every time.

4.4.4. Extra Pattern Detection

Extra features are sometimes produced inadvertently during the patterning of repeated structures. This is often seen in FIN and VIA layers. Conventional Die-to-Die detection methods are undependable because more than one die may have this issue. Anchor's computational system can reliably detect extra patterns because the chip design serves as the reliable reference, shown in *Figure 9*. This approach is also used to detect any extra feature on wafer caused by residue or fall-on particle.

4.4.5. Hole Analysis (Size Variation, Short, Missing)

Contacts and vias (i.e. *holes*) are printed by the billions on large logic devices, and by the hundreds of billions on each wafer at *each* hole layer. They play an essential role in the routing of electrical signals between layers. However, there can be considerable variation in the printing of holes. Variations can arise from natural process drift, from proximity effects of neighboring clusters, from randomness in the material and topography, from etch chamber control, etc.

Anchor's computational system monitors hole size and shape, detects various types of shrinkages and enlargements, and identifies missing holes, as shown in *Figure 10*. Moreover, it can automatically identify all holes in an image and analyze each one, leading to exceptionally thorough analysis.

4.5. Massive Pattern Fidelity Analysis

Pattern *fidelity* – not just *defectivity* – has always been of importance to the fab, but fidelity monitoring has been limited to low-frequency, time-consuming CD-SEM (Critical Dimension Scanning Electron Microscope) measurements [9,10]. CD-SEMs continue to play an important role in accurately

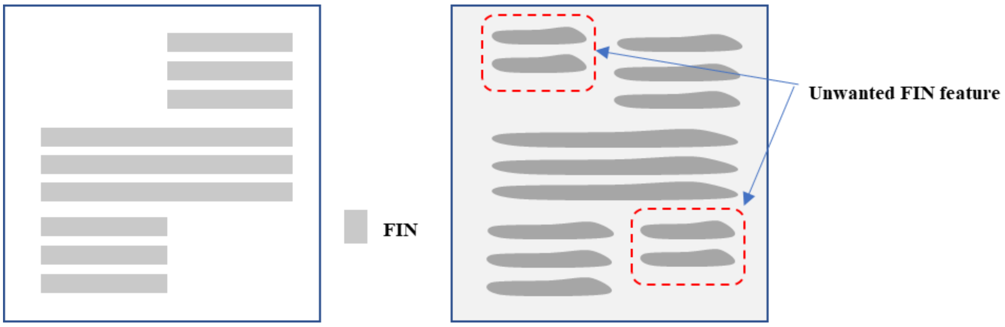


Figure 9. Target FIN feature (left) and Representative SEM image (right).

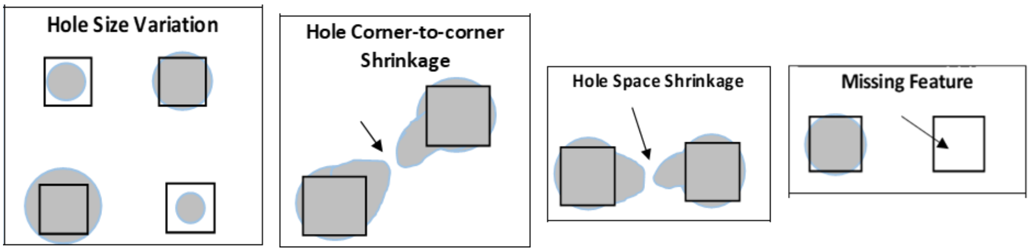


Figure 10. Hole defect detection.

measuring features from both orthogonal and oblique angles. These measurements are typically performed on preselected features on designated wafers.

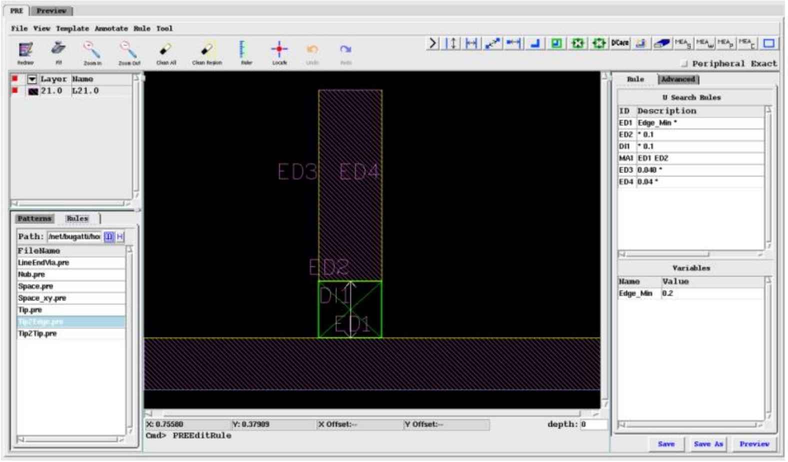
Anchor's computational system reimagines the concept and deploys it on a massive scale on all wafers and on all features for which there are Review SEM images. The line itself is monitored not only for the traditional concept of *defect*, but also for the concept of *pattern fidelity*, which is in effect a "CD"-type measurement, but without the same level of measurement accuracy as a calibrated CD-SEM measurement. As such, inline continuous massive pattern fidelity measurement supplements the conventional CD-SEM operation [11]. It has the potential to provide much earlier warnings of pending problems by tracking changes or trends in pattern fidelity *before* they become bona fide defects. At the leading technology nodes, even small changes in pattern fidelity lead to significant electrical parasitics or parametric issues. A *resistive via*, for instance, may be caused by a slightly narrow and therefore partially blocked via that can impact device timing characteristics, produce single bit failures in memory devices, and produce various other parametric problems. Line thinning, line edge roughness, corner rounding, corner-to-corner artifacts, etc. are all liable to cause parametric issues.

Anchor's computational system performs massive pattern fidelity analysis on each image, but does so in a pattern-centric manner that searches each aligned SEM image for all *patterns of interest* or POI, measures their printed dimensions, compares them against the reference design, and stores all results in the Printed Pattern Database.

Patterns-of-interest (POI) are based on one or more parametric search rules. POI can also be identified automatically from the Design Decomposition Database by searching for patterns with high risk scores. Here we provide an example based on parametric search. Tip-to-line is a common pattern-of-interest, in which the amount of gap between tip and line (among other parameters) may affect printability or *pattern fidelity*.

In the example shown in *Figure 11*, we use a graphical user interface (GUI) to create a tip-to-line rule. We specify several constraints such as the maximum width of the tip, the minimum length of the tip, and the maximum gap between tip and line. We want the rule to match tips whose widths are less than 100nm, whose lengths are at least 40nm, and the gap is at most 100nm.

This single rule will match all variations of tip-to-line where the tip width is less than 100nm, the tip length is greater than 40nm, and the gap is less than 100nm. When we run this rule against the two sample SEM images shown in *Figure 12*, we find a match where the reference tip-to-line gap (from design) is 64nm and another where the reference gap is 60nm. Once these patterns of interest (in blue) have been found, their *printed* sizes are measured either (a) from the image itself or (b) from the extracted contour. In the first example, the *measured*



U Search Rules

ID	Description
ED1	Edge_Min *
ED2	* 0.1
DI1	* 0.1
MA1	ED1 ED2
ED3	0.040 *
ED4	0.04 *

Variables

Name	Value
Edge_Min	0.2

Figure 11. Defining Tip-to-Line Rule with Parameters (Constraints).

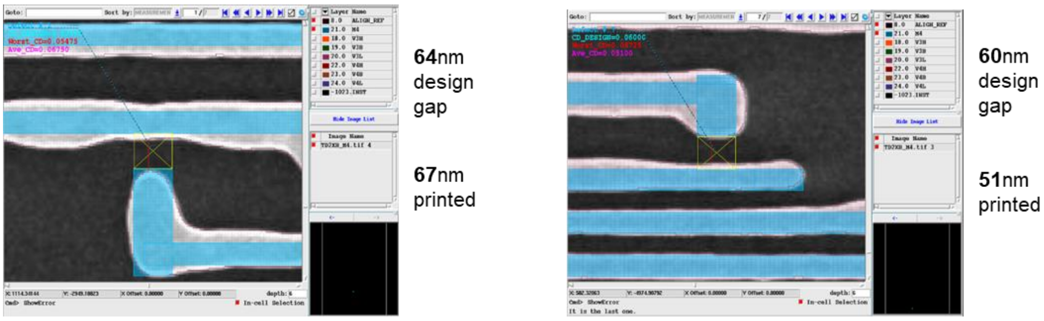


Figure 12. Two variations of tip-to-line matched by the parametric search rule.

value is 67nm, and in the second, the *measured* value is 51nm.

Because each SEM image is scanned from top to bottom, there might be tens or hundreds of matching patterns *on each image*. From a small set of sample images, we obtained the result as shown in *Figure 13*.

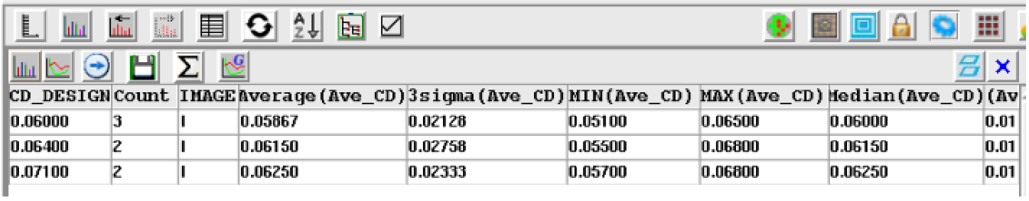
In column 1 we see that the tip-to-line rule found 3 variations of the pattern:

- Variation 1: Reference gap 60nm. Average of the printed gap was 58.67nm
- Variation 2: Reference gap 64nm. Average of the printed gap was 61.50nm
- Variation 3: Reference gap 71nm. Average of the printed gap was 62.50nm

Litho/OPC and process engineers can examine this table to study the effects of *gap size* on the overall fidelity of the printed pattern. They can ask questions such as *if the design or reference gap is reduced to X, how will that affect the printability of the pattern?* Similarly, *if the design or reference gap is enlarged to Y, how will that affect the printability of the pattern?* In other words, the effects of specific variations in the physical layout can be studied in a comprehensive manner.

This example also demonstrates the value of *speaking the universal language of patterns*. There are more examples shown in *Figure 14* that demonstrate the potential of SEM images to reveal detailed analysis of process variation. Anchor's computation system is like an "analog to digital" converter – it converts the rich information content of analog SEM images into concise digital design patterns while retaining all of the information associated with the analog print.

When we expand the example by using (a) multiple parametric search rules and (b) hundreds or thousands of SEM images, we obtain a deep understanding of the process and its limitations. For (a) each pattern type (e.g. tip-to-line, tip-to-tip, etc.) and (b) each *variation* of each pattern type (e.g. tip-to-line gaps of 60nm, 64nm, 71nm, etc.) we create a Box Plot that represents all of the measurements of that particular pattern variation. For instance, if we found and measured fifty tip-to-line patterns with intended gap of 60nm, we create a box plot that shows how close or how wide apart all of the



CD_DESIGN	Count	IMAGE	Average(Ave_CD)	3sigma(Ave_CD)	MIN(Ave_CD)	MAX(Ave_CD)	Median(Ave_CD)	(Av
0.06000	3	1	0.05867	0.02128	0.05100	0.06500	0.06000	0.01
0.06400	2	1	0.06150	0.02758	0.05500	0.06800	0.06150	0.01
0.07100	2	1	0.06250	0.02333	0.05700	0.06800	0.06250	0.01

Figure 13. Sample measurement results of tip-to-line.

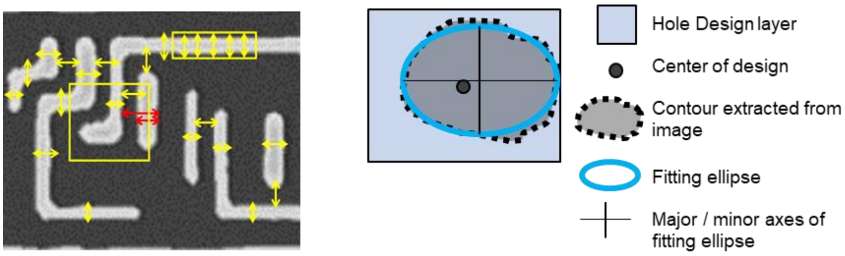

Hole Design layer
Center of design
Contour extracted from image
Fitting ellipse
Major / minor axes of fitting ellipse

Figure 14. Additional examples of various features being measured.

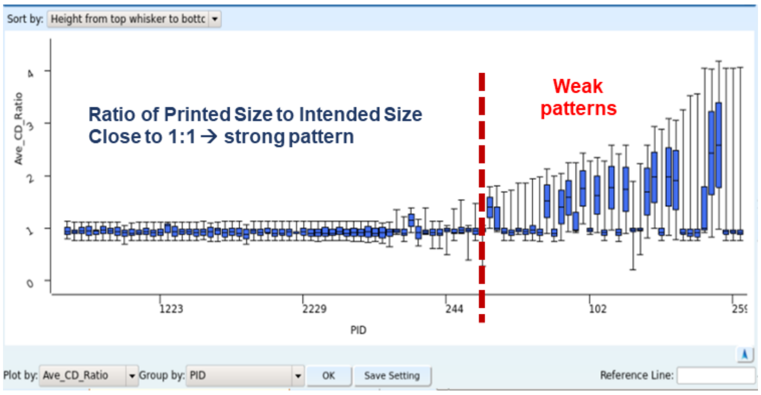


Figure 15. Box Plot of Various Patterns as a Function of the Ratio of Printed Pattern Measurement to Intended Pattern Measurement. Values closer to 1:1 indicate strong patterns.

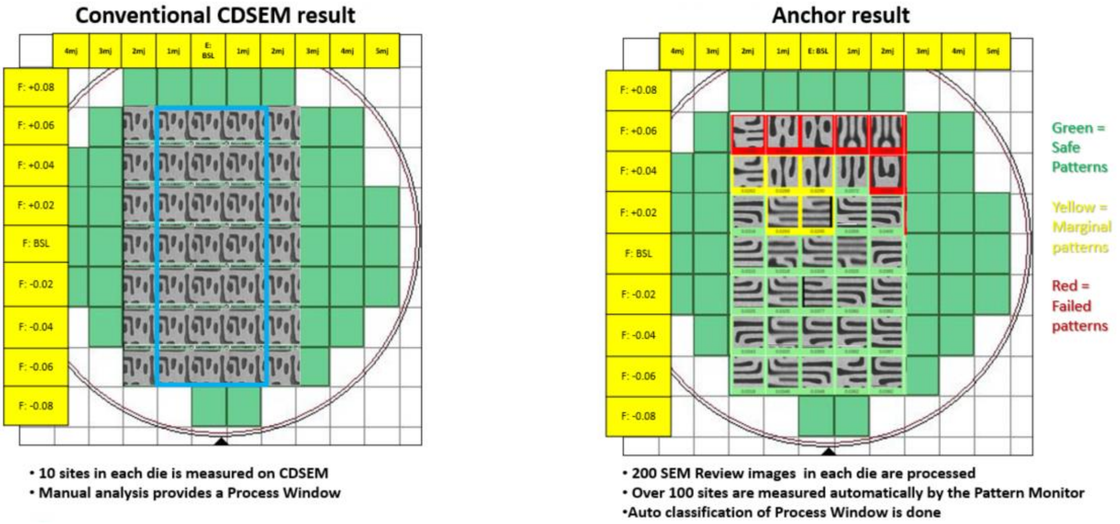
individual measurements were, and how much those measurements deviated from the reference of 60nm.

If we do the same for all patterns and their variations, we end up with a box plot as shown below. Each box represents the measurements of one *specific* pattern (e.g. tip-to-line with reference gap of 60nm). In this example we see numerous patterns.

Each box in a box plot shows several statistics about each *specific* pattern: the average and median values of all measurements, the range where most of the values are clustered, and outliers. It is a particularly effective way to identify weak and strong patterns, as shown in *Figure 15*. This particular box plot is based on the ratio of measured value to intended value. If the ratio is 1:1, it indicates a strong pattern because the *measured* values of all instances of that pattern matched the *intended* value. The more a box diverges from 1:1, the weaker the pattern. In this chart we see that the left half of patterns are printing well, with ratios close to 1, but the right half diverges significantly, indicating progressively weaker patterns. This automatically separates weak patterns from strong patterns, providing actionable information for root cause analysis.

Although this chart shows a large collection of patterns, we can track the behavior of individual patterns as well. Given a particular pattern A, we can:

- Build a box plot of its measurements by time and track the fidelity of pattern A day-by-day or week-by-week or before-and-after a mask or process revision.
- Build a box plot of its measurements sorted by process tool ID (e.g. scanner 1 or scanner 2, or etcher 1 or etcher 2, or chamber 1 or chamber 2) for (a) tool matching purposes, (b) identification of problematic tool or chamber, or (c) process drift monitoring.
- Build a box plot of its measurements by Focus / Exposure modulation on an PWQ or FEM



- 10 sites in each die is measured on CDSEM
- Manual analysis provides a Process Window

- 200 SEM Review images in each die are processed
- Over 100 sites are measured automatically by the Pattern Monitor
- Auto classification of Process Window is done

Figure 16. Conventional vs Anchor FEM Analysis [12].

wafer to study the subtle changes in the behavior of pattern A across F/E modulations. (See next section.)

4.6. Innovative FEM/PWQ Analysis and Process Window Determination

Lithography process window determination is a critical step in the setup and tuning of a scanner recipe. Two of the most significant recipe parameters are (a) focus offset and (b) exposure dose. Different patterns and different *neighborhoods* of those patterns are affected differently by focus and exposure settings, which are determined by exposing a reticle or mask using a series of focus and exposure *modulations* and analyzing the results of each modulation.

The conventional method of analyzing Focus/Exposure Modulations (FEM) is by performing a high-sensitivity wafer inspection followed by a large SEM Review in which tens of thousands of SEM images are captured and analyzed for the presence of *hard defects*. The conventional method does not take *pattern fidelity* into account and therefore cannot track or report the subtle deviations that occur on each pattern across each modulation. Subtle deviations – pattern *fidelity* variations – are playing an increasingly significant role in parametric yield loss. Establishing a lithography process window that takes into account pattern fidelity (not just pattern defectivity) leads to a more robust result [12-15]. A side-by-side comparison of the process window map obtained by conventional method and by Anchor's method is shown in *Figure 16*. Anchor's computational system redefines and reinvents FEM/PWQ analysis in the following ways:

- The computational system checks every SEM image for the presence of die-to-database defects. Some of these defects are not detectable using conventional die-to-die or die-to-golden die techniques. Multiple defects can be detected on a single image.
- The computational system *measures* every feature of interest in every SEM image (massive metrology) to generate pattern uniformity statistics for each pattern of interest. This enables pattern *fidelity* analysis.
- The computational system tracks the uniformity of like patterns across each modulation to generate Bossung Curves automatically for hundreds or thousands or tens of thousands of patterns. These Bossung Curves supplement – not replace – conventional CD-SEM analysis because accuracy of measurements from Review SEM is limited. Nevertheless, these Bossung Curves are produced more quickly and cover a significantly wider set of patterns. They provide valuable early feedback.

The combination of (a) better defect detection, (b) pattern uniformity/fidelity analysis, and (c) generation of Bossung Curves for a wide set of patterns results in the reinvention of PWQ/FEM analysis.

4.7. Risk Assessment of New Tape-outs

Historically, it has been difficult to comprehensively assess the yield risk of a new incoming device. This requires the device to be searched for known weak patterns in order for corrective action to be taken by Litho/OPC teams before the mask is made.

The Design Decomposition Database, in which all patterns of interest are ranked by a machine learning model built from real printed wafer images, enables comprehensive full-chip pattern risk assessment for new tape-outs. The new tape-out is decomposed into constituent patterns that are both (a) cross-referenced with existing patterns in the Design Decomposition Database and (b) assigned risk scores directly by the trained machine learning model. The new tape-out, therefore, is systematically evaluated for potential risk, and corrective action can be taken well in advance of printing the (expensive) masks.

4.8. Large Scale in-wafer OPC Verification

OPC simulations are standard practice in most fabs. They are based on complex and finely tuned models of the lithography column, and often take hours or days to run on a large cluster of computing nodes (servers). OPC simulations produce a report that grades the lithography risk of each pattern (including the neighborhood in which the pattern lies). Some patterns are clearly marked “weak”, others are “borderline weak”, and others might be “unknown”.

An OPC result is a set of patterns and their risk assessments. But these patterns are very difficult to verify in the fab because once the reticle is printed, a *digital-to-analog* conversion has taken place. SEM images are analog bitmaps, and these images cannot be compared directly with the OPC simulation results. Instead, images (analog) must be converted back to patterns (digital). This *analog-to-digital* conversion is once again the basis for Anchor’s *pattern-centric* computational system. It allows thousands or millions of SEM images to be converted back into digital (pattern) representations that can finally be compared with OPC simulation results in a comprehensive manner to assess the validity of the OPC model. Specifically, we can answer such questions as:

- If OPC simulation predicted a weak pattern, was that pattern *actually* weak? If we examine the box plot of that pattern, we can answer the question immediately.
- If OPC simulation predicted a strong pattern, was that pattern actually strong?
- Did OPC simulation fail to predict a weak pattern (alpha risk)? If so, results from Anchor’s computational system can be used to fine-tune the OPC model.

5. Conclusion

Anchor has developed a pattern-centric computational system for the fab that rests on the three pillars of (a) printed pattern database, (b) design decomposition database, and (c) machine learning. These pillars extract significantly richer information from the analog or printed wafer domain, convert it into the digital or pattern-based domain, and enable wide-ranging applications for yield learning, defect discovery, line monitoring, and design-process co-optimization. The computational system is vendor-neutral and has been adopted at multiple Tier-1 and Tier-2 fabs around the world.

References

[1] Khurram Zafar, Chenmin Hu, Ye Chen, Yue Ma, Chingyun Hsiang, Justin Chen, Raymond Xu, Abhishek Vikram, Ping Zhang, "Pattern weakness and strength detection and tracking during a semiconductor device fabrication process", US Patents #9,846,934 (2017), #10,062,160 (2018).

[2] Khurram Zafar, Chenmin Hu, Ye Chen, Yue Ma, Chingyun Hsiang, Justin Chen, Raymond Xu, Abhishek Vikram, Ping Zhang, "Pattern weakness and strength detection and tracking during a semiconductor device fabrication process", Taiwan Patents #I608427(2017), #I634485 (2018).

[3] Chenmin Hu, Khurram Zafar, Chen Ye, Ma Yue, Lv Rong, Justin Chen, Abhishek Vikram, Yuan Xu, Ping Zhang, "Pattern Centric Process Control", US Patent 10,546,085 (2020).

[4] Eric Guo, Shirley Zhao, Skin Zhang, Sandy Qian, Guojie Cheng, Abhishek Vikram, Ling Li, Ye Chen, Chingyun Hsiang, Gary Zhang, and Bo Su, "Simulation based mask defect repair verification and disposition”, Proc. SPIE 7488, Photomask Technology 2009, 74880G, 2009.

[5] Eric Guo, Irene Shi, Blade Gao, Nancy Fan, Guojie Cheng, Li Ling, Ke Zhou, Gary Zhang, Ye Chen, Chingyun Hsiang, and Bo Su, "Simulation based mask defect printability verification and disposition, part II", Proc. SPIE 8166, Photomask Technology 2011, 81662D, 2011.

[6] Li-Fu Chang, Chang-Il Choi, Guojie Cheng, Abhishek Vikram, Gary Zhang, and Bo Su, "Detection of OPC conflict edges through MEEF analysis", Proc. SPIE 7641, Design for Manufacturability through Design-Process Integration IV, 764111, 2010.

[7] Abhishek Vikram, Kuan Lin, Janay Camp, Sumanth Kini, Frank Jin, Vinod Venkatesan, “Inspection of high-aspect ratio layers at sub 20nm node”, Metrology, Inspection, and Process Control for Microlithography XXVII, Proc. of SPIE Vol. 8681, 86811Q, 2013.

[8] Jing Zhang, Qingxiu Xu, Xin Zhang, Xing Zhao, Jay Ning, Guojie Cheng, Shijie Chen, Gary Zhang, Abhishek Vikram, Bo Su, "Yield impacting systematic defects search and management", Design for Manufacturability

through Design-Process Integration VI, Proc. of SPIE Vol. 8327, 832716, 2012.

[9] Gyun Yoo, Jungchan Kim, Taehyeong Lee, Areum Jung, Hyunjo Yang, Donggyu Yim, Sungki Park, Kotaro Maruyama, Masahiro Yamamoto, Abhishek Vikram, Sangho Park, "OPC verification and hotspot management for yield enhancement through layout analysis", Metrology, Inspection, and Process Control for Microlithography XXV, Proc. of SPIE Vol. 7971, 79710H, 2011.

[10] Taehyeong Lee, Hyunjo Yang, Jungchan Kim, Areum Jung, Gyun Yoo, Donggyu Yim, Sungki Park, Akio Ishikawa, Masahiro Yamamoto, Abhishek Vikram, "Hot spot management through design-based metrology: measurement and filtering", Proc. SPIE. Vol. 7520, 75201U, 2009.

[11] Sicong Wang, Jian Mi, Abhishek Vikram, Gao Xu, Guojie Cheng, Liming Zhang, Pan Liu, "Novel pattern-centric solution for high performance 3D NAND VIA dishing metrology", Design-Process-Technology Co-optimization for Manufacturability XIII, SPIE Vol. 10962, 1096217, 2019.

[12] Ming Tian, Yu Zhang, Tiapeng Guan, Jianghua Leng, Baojun Zhao, Lei Yan, Wei Hua, Abhishek Vikram, Guojie Chen, Hui Wang, Gary Zhang, Wenkui Liao, "Critical Defect Detection, Monitoring and Fix through Process Integration Engineering by Using D2DB Pattern Monitor Solution", Design-Process-Technology Co-optimization for Manufacturability XIII, SPIE Vol. 10962, 109620L, 2019.

[13] Lijun Chen, Jun Zhu, Xuedong Fan, Haichang Zheng, Xiaolong Wang, Yancong Ge, Yu Zhang, Abhishek Vikram, Guojie Cheng, Hui Wang, Qing Zhang, Wenkui Liao, "An Advanced and Efficient Methodology for Process Setup and Monitoring by Using Process Stability Diagnosis in Computational Lithography", Design-Process-Technology Co-optimization for Manufacturability XIV, Proc. SPIE. 11328, 2020.

[14] Qian Xie, Panneerselvam Venkatachalam, Julie Lee, Zhijin Chen, Khurram Zafar, "Design guided data analysis for summarizing systematic pattern defects and process window", Proc. SPIE. 9778, Metrology, Inspection, and Process Control for Microlithography XXX Proceedings Article, 2016.

[15] Qian Xie, Panneerselvam Venkatachalam, Julie Lee, Zhijin Chen, Khurram Zafar, "Precise design-based defect characterization and root cause analysis", Proc. SPIE. 10145, Metrology, Inspection, and Process Control for Microlithography XXXI Proceedings Article, 2017.

Photography & Biography

Chenmin Hu is the founder and CEO of Anchor Semiconductor Inc. He has over 40 years of industry experience and has previously worked with Synopsys and Valid Logic. He holds Ph.D. degree in Electrical Engineering.

Khurram Zafar is Vice President of Applications Engineering and Technical Marketing at Anchor Semiconductor. He has over 30 years of industry experience and has previously worked with KLA and Texas Instruments. He holds B.S. degree in Electrical Engineering.

Abhishek Vikram is Director of Applications Engineering and Technical Marketing at Anchor Semiconductor. He has over 19 years industry experience and has previously worked with KLA and GlobalFoundries. He holds Ph.D. degree in Electrical Engineering.

Geoffrey Ying is Vice President of Business Development and Product Marketing at Anchor Semiconductor. He has over 30 years of industry experience and has previously worked with Cadence and Synopsys. He holds M.S. degree in Electrical Engineering and MBA.

A Global Cutting-Edge Semiconductor Technology Reporting Journal

Journal of Microelectronic Manufacturing

Available on www.jommpublish.org

Included by the Chinese Science and Technology Journal (CSTJ) Database