SISPAD 2022

INTERNATIONAL CONFERENCE ON SIMULATION OF SEMICONDUCTOR PROCESSES AND DEVICES

Conference Abstract Booklet

UNIVERSIDAD
DE GRANADA

2022 International Conference on Simulations of Semiconductor Processes and Devices

September 6-8th, Granada (Spain)

Welcome to SISPAD 2022

Dear Colleague,

On behalf of the University of Granada and the Organizing Committee, I warmly welcome you to the 27th International Conference on Simulations of Semiconductor Processes and Devices. It has been a difficult period with the worldwide pandemic. The last two editions were very complicated, and I would like to publicly recognize the efforts and generosity of the Organizing Committees of SISPAD-2020 (Osaka, Japan) and SISPAD-2021 (Dallas, TX, USA) for keeping the conference even in so difficult circumstances. We're all in debt to them. Thank you !!

We would also like to make a special mention to the Steering Committee of SISPAD and its President Prof. Juergen Lorentz, for their support, professionality, and their valuable advice during the organization of the conference.

I am very grateful that so many of you have been able to overcome the difficulties and make it to Granada. This run of SISPAD 2022 has once again attracted an extremely high standard of participants and I am sure that this will be an extraordinary meeting. Our delegate list includes contributions from twenty-two countries, from many leading companies, Universities and Research Labs. So I anticipate plenty of stimulating discussions. I also hope that SISPAD 2022 will promote new collaborations between industry and academia which will further advance the science of Semiconductor simulations.

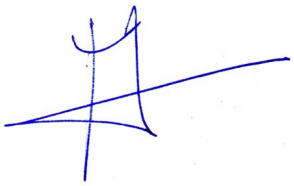
The success of a meeting like SISPAD depends, of course, on the quality of the presentations but even more important is the active participation of the other assembled experts. So, I encourage everyone to take part in the discussions and debates that go on both during the scientific sessions and during the breaks. We have tried to put all the ingredients so the conference becomes a familiar and comfortable event, and you feel at home.

We have not spared any efforts to allow the SISPAD conference to recover the excellence that it used to have in the events previous to the pandemic, not only in the scientifical sphere but also in the social one. And, here we have counted with the inestimable help of the city of Granada. Granada has an incomparable historical, cultural and social heritage. For thousands of years, Granada has served as a place of convivence for millenary civilizations, Jews, Christians and Muslims. Granada and its Alhambra represent an incomparable frame of mixing of cultures, and peaceful coexistence. And all these have deeply marked the actual life and spirit of an old, but at the same time mod-

ern city and its open-mind population.

Granada is also the seat of one of the older Universities in Spain and in Europe. Founded by Emperor Charles V in 1531, the University of Granada with almost 500 years of history, is today one of the more active Universities in Spain (in the top 3) and it is ranked in the top 200 of all Universities Worldwide.

I would like to thank you all – presenters, audience, exhibitors, organizers and the SISPAD committees (Steering Committee and Technical Programme Committee) - for your participation in this conference. I hope you have an interesting, informative, enjoyable, and overall, unforgettable stay at Granada.



F. Gamiz
SISPAD 2022 Conference Chair

Contents

About Electron Transport and Spin Control in Semiconductor Devices

Siegfried Selberherr[1] and Viktor Sverdlov[1,2]
[1]Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, A-1040 Wien, Austria
[2]Christian Doppler Laboratory for Nonvolatile Magnetoresistive Memory and Logic at the Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, A-1040 Wien, Austria
e-mail: {Selberherr|Sverdlov}@iue.tuwien.ac.at

***Abstract*—The electron charge and its response to the electrostatic field resulting in electrical charge redistributions and currents is at the heart of the complimentary metal-oxide-semiconductor (CMOS) transistors' functionality. As the scaling of CMOS-based technology displays signs of an imminent saturation, an introduction of novel non-conventional computational degrees of freedom to sustain the path of energy efficient computing at reduced costs becomes paramount. Employing the second intrinsic electron characteristics – the electron spin – offers additional functionality to electron devices and may further boost the performance of integrated circuits. Apart from forming a qubit, the electron spin is promising for digital applications. SpinFETs and SpinMOSFETs are devices using the spin polarization, with nonvolatility introduced through the relative magnetization orientation between the ferromagnetic source and drain. Several fundamental problems including the spin injection from metal ferromagnets to a semiconductor, spin propagation and relaxation, as well as spin manipulation by the gate voltage were successfully demonstrated recently providing a vision that these devices may appear in significant numbers on the market in the non-distant future.**

On the memory side, the nonvolatile CMOS-compatible spin-transfer torque (STT) and the spin-orbit torque (SOT) magnetoresistive random access memories (MRAMs) are already competing with flash memory. MRAM is expected to replace DRAM and SRAM in embedded applications. MRAM possesses a simple structure, long retention time, high endurance, and it is fast. A combination of nonvolatile elements with CMOS circuitry allows to shift the data processing into the nonvolatile segment paving the way for a novel low power computational paradigm based on logic-in-memory and in-memory computing architectures.

To model MRAM, an appropriate extension of the spin and charge transport equations to multi-layered structures consisting of normal and ferromagnetic metal layers separated by tunnel barriers is required. The charge current is accurately reproduced by a low conductivity locally depending on the angle between the magnetization vectors, while for the spin current proper boundary conditions at the tunnel barrier interfaces are needed. We apply our approach to model the magnetization dynamics in ultra-scaled MRAM cells with composite elongated free layers.

Keywords— charge and spin currents, TCAD, digital spintronics, SpinFET, SpinMOSFET, STT-MRAM, SOT-MRAM.

I. Introduction

Continuous miniaturization of metal-oxide-semiconductor field effect transistors (MOSFETs) is one of the main drivers ensuring the impressive increase in speed, performance, density, and complexity of modern integrated circuits. Numerous outstanding technological challenges have been resolved on the exciting journey of continuous transistor downscaling. At all stages, accurate TCAD tools were paramount to predict the device functionalities, to optimize the parameters, and to obtain the best performance. To describe the electron transport, the drift-diffusion model enjoyed a remarkable success at earlier stages of MOSFETs modeling due to its relative simplicity, numerical robustness, and the ability to perform two- and three-dimensional simulations on large unstructured meshes [1].

Since the 90nm technology node, strain as a MOSFET channel mobility booster was introduced [2]. This allowed improving the on-current while maintaining other characteristics including the off-current unchanged. To improve the on-current in *n*-silicon channels, tensile capping layers were used. In *p*-channels compressive capping layers were complemented by epitaxially re-grown SiGe source and drain enhancing strain [2]. The simulation tools incorporated corresponding strain-enhanced mobility models which were carefully tailored with the help of accurate band structure calculations of strained Si [3].

To proceed to the 45nm CMOS technology node, the electrostatic channel control was improved by replacing the native SiO_2 gate oxide with a high permittivity HfO_2 dielectric and metal gates [4]. The use of HfO_2 allowed to maintain the physical gate oxide sufficiently thick to suppress parasitic tunneling leakage currents. However, to preserve a good electrostatic channel control at the 22nm technology node, a new architecture of three-dimensional tri-gate transistors was required [4]. Importantly, the doping of three-dimensional fins became unnecessary as the electron confinement in the channel is geometrical and is not due to the depletion. This allowed to make the fins taller and thinner, which, in combination with strain and high-k dielectrics/metal gates allowed to scale the technology down to 14nm [5] and 10nm [6]. By introducing pitch splitting, self-aligned patterning, and EUV lithography the transistor miniaturization was continued down to 7nm feature size [7]. Using full-fledged EUV lithography and high mobility channel FinFETs, the 5nm technology node provides ~1.8x improvement in logic density, 15% speed gain, and 30% power reduction compared to its previous generation [8]. For upcoming technology nodes, gate-all-around devices are optimal candidates. To sustain the high on-current through the transistor, the channel in such devices should contain several nanosheets surrounded by the gates [9].

To model such devices, quantum effects including a confinement at least in the direction orthogonal to transport must be included [10]. When modeling the transport in ultra-scaled fins with only a few subbands occupied, a full subband method [11] must be applied.

For sufficiently long transistor channels the carrier motion in transport direction can often be treated semi-classically, and the use of transport models based on a set of coupled subband Boltzmann transport equations is required [12]. In ultra-scaled MOSFETs, however, the current becomes close to ballistic and is determined by several transversal propagating modes in the channel. In this case the channel conductance ceases to depend on the channel length. It then follows that in order to maintain the same current through the channel, the supply voltage must be fixed. It does not decrease with the channel length regardless of its reduction [13]. This results in an approximately constant power dissipation of a single MOSFET, so the generated heat per area increases rapidly with increasing transistor density, which puts a foreseeable limitation on scaling.

II. Electron Spin as a Computational Degree

As the downscaling of charge-based CMOS devices is approaching its saturation, the electron spin attracts much attention as a suitable candidate for complementing or even replacing the charge degree of freedom in future microelectronic devices [14], [15], [16]. The electron spin can point in any direction on a Bloch sphere. A single spin is suitable to build a qubit as opposed to a bit of binary information. Two or more qubits can be in an entangled state. This state cannot be represented as a direct product of independent qubits. The entangled state is therefore characterized by quantum mechanical correlations which cannot be described by classical mechanics. Therefore, using qubits for quantum information processing guarantees a superiority of a quantum computer over a classical one for certain applications. Single-spin qubits in semiconductors are realized with impurities [17] or quantum dots [18]. Single-spin qubits in diamond and SiC demonstrate a spin coherence time of several microseconds even at room temperature [19]. The quantum dot technology was recently shown to have a potential towards utilization in a large-scale quantum computer. It allows to process three-dimensional silicon quantum chips, in which a precise interlayer alignment is achieved and, therefore, a highly accurate manipulation and detection of the spin states is possible [20]. A successful implementation of a quantum computer based on spins in quantum dots requires the capability of efficient spin initiation, coherent manipulation, and reliable read-out [21]. An unprecedented advantage in these fields has been demonstrated with the successful realization of a two-qubit quantum gate [22] as well as of a quantum processor with a fidelity exceeding 99% [23]. The pressing challenge is to proceed to a larger computational network of qubits. Employing quantum dots is an attractive option due to their similarity to scaled transistors giving the perspective to integrate spin qubit devices on a 300mm wafer [24].

Spins of impurities in solids represent another attractive option to create qubits. Spins of impurities can be combined with nearby nuclear spins. Recently, a three-qubit quantum processor [25] as well as a ten-qubit solid-state spin register with a long-time stable quantum memory up to one minute [26] were demonstrated. However, the technology requires a controllable creation of defects with long-living spin states, which represents a significant challenge in modern semiconductor technology.

The stationary spin state is characterized by one of the two possible projections on a given axis. Therefore, if the axis orientation is given, for example, by means of a magnetic field or a magnetization direction, the two states of spin projections can be used for classical data processing. Therefore, apart from quantum applications, the electron spin offers a great opportunity to design devices with superior characteristics for digital applications. Modeling approaches and status of spin augmented electron switches as well as a spin current enabled magnetoresistive random access memories (MRAM) are discussed next.

III. SpinFETs and SpinMOSFETs

In addition to the electron charge, a spin field-effect transistor (SpinFET) [27] employs the spin polarization to enrich the transistor performance. A SpinFET is composed of a semiconducting channel region sandwiched between the two ferromagnetic contacts, source and drain. The ferromagnetic source contact injects spin-polarized electrons in the semiconductor region. As the electron reflection depends on the ferromagnetic drain orientation relative to the current polarization, the electrons with their spins aligned to the drain contact magnetization can leave the channel more easily through the drain contact, thus contributing most to the current. Therefore, the electric current depends on the drain magnetization orientation relative to the spin polarization of the current close to the drain. Thus, the drain magnetization can be used to detect the spin polarization of the electric current. In the case of parallel or anti-parallel orientations of the source and drain the on-currents through the transistor are different. This opens a path to a reconfigurable logic [28], when the same device possesses different characteristics depending on the source-drain relative magnetization orientation. The spin-related signal is therefore the on-current difference in the parallel and anti-parallel configuration of the source and drain magnetizations. Experimentally, a large absolute (but not relative) difference between the two values of the on-currents for parallel and anti-parallel source-drain orientations was recently observed in a Si MOSFET at room temperature [29].

In contrast to the electron charge, the electron spin polarization injected into the channel with the charge current relaxes to its equilibrium zero value while propagating through the channel. In [29] the distance between the source and drain was several micrometers. It implies that, if the spin polarized current was injected into the Si channel through the source, it remains spin-polarized at the drain after the electrons flow through the channel. This finding signifies that the spin diffusion length in silicon at room temperature is long and sufficient for spin-based digital applications. The spin relaxation can be further significantly suppressed by using stress [30], the technique commonly applied to boost the carrier mobility in Si [2].

However, there is one caveat: Due to the spin conductivity mismatch [31] between the ferromagnetic metal contact and the semiconductor, a spin polarization cannot be dynamically injected into the semiconductor with the current. As silicon is a non-magnetic semiconductor, the densities of states for spin-up and spin-down electrons are equal. When the current flows through the interface between a ferromagnetic metal and a semiconductor, the supply of both spin-up and spin-down electrons in the semiconductor is overwhelmingly high. Both states with spin-up and spin-down are equally populated. The current running through the semiconductor remains unpolarized.

After the spin-impedance mismatch [31] problem between the ferromagnetic metal and a semiconducting channel had been understood, a solution to inject spin-polarized current was soon proposed [32]. By inserting a tunnel barrier in the ferromagnet-semiconductor contact the flux of the electrons is reduced to such an extent that the densities of state-up/down in the semiconducting channel are populated proportionally to the majority/minority spins in the ferromagnet. The tunnel barrier must neither be too thick nor too thin. In the famous SpinMOSFET demonstration [29], MgO tunnel barriers guaranteed a successful spin injection and detection with ferromagnetic source and drain contacts.

The idea to manipulate the electron spins by the gate voltage proposed in [27] has inaugurated a new field of spin-based electronics, i.e., spintronics. However, the demonstrated device [29] is not a real SpinFET [27] as the (back)gate only serves to turn the current on or off. The proposal [27] constitutes an all-electric, not magnetic, way to controllably manipulate spins by an effective spin-orbit field. As the strength of the spin-orbit field depends on the electric field at the interface [33], which is modulated by the gate voltage, it makes the spins to precess at a certain gate voltage-dependent angle, while the spins are propagating in the channel under the gate. The fact that the spin-orbit field depends on the electron momentum guarantees that all ballistically propagating electrons acquire the same spin rotation [27] at the end of the channel. To demonstrate a SpinFET one should have a semiconductor with a strong spin-orbit interaction, which is beneficiary for manipulating the spins. The spin-orbit interaction is several orders of magnitude larger in III-V semiconductors compared to the one in Si making them good candidates to build SpinFETs. However, a strong spin-orbit interaction also results in a strong spin relaxation. Therefore, a SpinFET made with III-V channels can only operate at cryogenic temperatures. For a demonstration of the gate-induced spin precession it is necessary to avoid ferromagnets as they can cause stray fields resulting in spurious current oscillations. An elegant solution to the problem of injecting spin-polarized electrons without using ferromagnetic contacts was recently [34] proposed. The effective spin-orbit field [33] guarantees that the moving electrons become automatically spin-polarized. Injecting spins through a point contact biased with a normal electric field created by additional gates results in spin-polarized injection currents. Using this injection technique, the current modulation by means of the gate–voltage-dependent spin-orbit field was confirmed [34]. This is the first successful demonstration of a SpinFET at cryogenic temperatures [34] 25 years after it was predicted [27].

Although many fundamental problems and technological challenges have been resolved and a SpinFET and a SpinMOSFET were successfully demonstrated, both devices still rely on the charge current to transfer the spin and, therefore, are prone to the same limitations as CMOS FETs. In addition, the absence of an efficient and purely electrical spin injection scheme into a semiconducting channel results in a small relative difference between the on-current in the parallel and anti-parallel source-drain configuration. New innovative solutions for purely electrical efficient spin injection and manipulation are urgently needed. Exploring special properties of novel two-dimensional materials is one possible avenue: Recently, it was shown that the in-plane electric field in a sheet of tungsten diselenide bonded on a sheet of bilayer graphene is capable of changing the polarization of the spin current up to room temperature [35]. An alternative path is to explore devices which already possess a large magnetoresistance.

IV. Emerging Magnetoresistive Memories

A magnetic tunnel junction (MTJ) is a sandwich made of two ferromagnetic contacts separated by a tunnel junction. The tunnel magnetoresistance ratio in MTJs can reach several hundred percent at room temperature [36]. Due to the large difference in the resistances of the parallel and anti-parallel MTJ configurations, MTJs are suitable for storing the binary data. The typical resistance of an MTJ is similar to that of a MOSFET. It makes MTJ-based memories electrically compatible with CMOS circuitry without extra amplifiers to convert the spin (magnetization) degree into charge. The spin-transfer torque effect [37], [38] has been shown to be perfectly suitable for the electrical data writing. Indeed, the magnetization of a free layer (FL) can be changed between the two orientations parallel and antiparallel to the magnetization of the fixed layer by passing an electric current through an MTJ.

STT-MRAM is considered as a perfect candidate for future universal memory. STT-MRAM is fast (10ns), it possesses high endurance (10^{12}), and it has a simple structure. It is compatible with CMOS technology and can be straightforwardly embedded in circuits [39]. STT-MRAM solutions compatible with 22nm FinFET [40] and 16nm FD SOI [41] technologies are available. An MRAM cell consists of several layers, including CoFeB ferromagnetic reference and free layers separated by an MgO tunnel barrier. To increase the perpendicular magnetic anisotropy, the FL, typically composed of two CoFeB layers and a thin metal buffer, is interfaced with the second MgO layer [42]. Introducing more MgO layers [43] and elongating the FL allows to boost the perpendicular anisotropy even further, while reducing the FL diameter [44]. For an accurate design of ultra-scaled MRAM cells it is paramount to properly incorporate ferromagnetic layers separated by normal metal buffers MgO barriers in between [45]. Incorporating torques acting on the textured magnetization in elongated ultra-scaled FLs [45] is important The coupled spin and charge transport approach [45] coupled with temperature [46], [47] was demonstrated to be applicable [48] to double reference layer STT-MRAM [49] as well as it can be generalized to describe spin-orbit torques (SOT) [50] in SOT-MRAM.

V. Conclusion

A successful adoption of the electron spin in microelectronics can potentially revolutionize data processing by, e.g., introducing spin-based qubits and the use of the electron spin in digital switches. Both fields are still in the fundamental research phase, just the emerging magnetoresistive memories already employ the spin currents for their operation. MRAM is fast nonvolatile, and CMOS-compatible. Developing logic-in-memory architectures suitable for in-memory processing will inevitably improve ultralow-power electronics, Big Data analysis, automotive electronics, and the Internet of Things. The electron spin can also be expected to have an enormous impact on neuromorphic computing and artificial intelligence of things.

ACKNOWLEDGMENT

Financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

REFERENCES

[1] S. Selberherr: Analysis and Simulation of Semiconductor Devices (Springer 1984).

[2] S.E. Thompson, M. Armstrong, C. Auth et al.: A Logic Nanotechnology Featuring Strained-Silicon, IEEE Electron Device Lett. 25, 191-193 (2004).

[3] V Sverdlov: Strain-Induced Effects in Advanced MOSFETs; S.Selberherr (ed) (Springer 2011).

[4] K. Mistry, C. Allen, C. Auth et al.: A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-Free Packaging, in Proc. IEDM Conf., 2007, pp.247-250.

[5] M. Bohr: The Evolution of Scaling from the Homogeneous Era to the Heterogeneous Era, in Proc. IEDM Conf., 2011, pp.1.1.1-1.1.6.

[6] S. Natarajan, M. Agostinelli, S. Akbar et al.: A 14nm Logic Technology Featuring 2nd-Generation FinFET, Air-Gapped Interconnects, Self-Aligned Double Patterning and a 0.0588 μm^2 SRAM Cell Size, in Proc. IEDM Conf., 2014, pp.3.7.1-3.7.3.

[7] C. Auth, A. Aliyarukunju, M. Asoro et al.: A 10nm High Performance and Low-Power CMOS Technology Featuring 3rd Generation FinFET Transistors, Self-Aligned Quad Patterning, Contact over Active Gate and Cobalt Local Interconnects, in Proc. IEDM Conf., 2017, pp.29.1.1-29.1.4.

[8] R. Xie, P. Montanini, K. Akarvardar et al.: A 7nm FinFET Technology Featuring EUV Patterning and Dual Strained High Mobility Channels, in Proc. IEDM Conf., 2016, pp.2.7.1-2.7.4.

[9] G.Yeap, S.S. Lin, Y.M. Chen et al.: 5nm CMOS Production Technology Platform Featuring Full-Fledged EUV, and High Mobility Channel FinFETs with Densest 0.021μm^2 SRAM Cells for Mobile SoC and High Performance Computing Applications, in Proc. IEDM Conf., 2019, pp.36.7.-36.7.4.

[10] M. Nedjalkov, I. Dimov, S. Selberherr: Stochastic Approaches to Electron Transport in Micro- and Nanostructures (Birkhäuser Basel, 2021).

[11] D. Esseni, P. Palestri, L. Selmi: Nanoscale MOS Transistors: Semi-Classical Transport and Applications (Cambridge University Press 2011).

[12] N. Loubet, T. Hook, P. Montanini et al.: Stacked Nanosheet Gate-All-Around Transistor to Enable Scaling Beyond FinFET, Proc. Symp. VLSI Technol., 2017, pp.T230-T231.

[13] M. Lundstrom: Fundamentals of Nanotransistors (World Scientific 2017).

[14] M. Johnson. Magnetoelectronics (Elsevier 2004).

[15] J. Fabian, A. Matos-Abiaguea, C. Ertler et al.: Semiconductor Spintronics, Acta Phys. Slovaca 57, 565-907 (2007).

[16] R. Jansen: Silicon Spintronics, Nature Materials 11, 400-408 (2012).

[17] J.J. Pla, K.Y. Tan, J.P. Dehollain et al.: A Single-Atom Electron Spin Qubit in Silicon, Nature Letters 489, 541-545 (2012).

[18] J.J.L. Morton, D.R. McCamey, M.A. Eriksson, S.A. Lyon: Embracing the Quantum Limit in Silicon Computing, Nature 467,687-691 (2011).

[19] Q.Li, J.-F. Wang, F.-F. Yan et al.: Room-Temperature Coherent Manipulation of Single-Spin Qubits in Silicon Carbide with a High Readout Contrast, National Sci. Rev. 9, nwab122 (2022).

[20] M. Koch, J.G. Keizer, P. Pakkiam et al.: Spin Read-Out in Atomic Qubits in an All-Epitaxial Three-Dimensional Transistor, Nature Nanotechnology 14, 137–140 (2019).

[21] D. Loss, D.P. DiVincenzo: Quantum Computation with Quantum Dots, Phys. Rev. A 57, 120-126 (1998).

[22] M. Veldhorst, C.H. Yang, J.C.C. Hwang et al.: A Two-Qubit Logic Gate in Silicon, Nature 526, 410-414 (2015).

[23] A.R. Mills, C.R. Guinn, M.J. Gullan et al.: Two-Qubit Silicon Quantum Processor with Operation Fidelity Exceeding 99%, Science Advances 8, eabn5130 (2022).

[24] A.M.J. Zwerver, T.Krähenmann, T.F. Watson et al., Qubits Made by Advanced Semiconductor Manufacturing, Nature Electronics 5, 184-190 (2022).

[25] M.T. Mądzik, S. Asaad, A. Youssry et al., Precision Tomography of a Three-Qubit Donor Quantum Processor in Silicon, Nature 601, 348-253 (2022).

[26] C.E. Bradley, J. Randall, M.H. Abobeih et al., A Ten-Qubit Solid-State Spin Register with Quantum Memory up to One Minute, Phys.Rev. 9, 031045 (2019).

[27] S. Datta, B. Das: Electronic Analog of the Electro-Optic Modulator, Appl. Phys. Lett. 56 (7), 665-667 (1990).

[28] S. Sugahara, J. Nitta: Spin-Transistor Electronics: An Overview and Outlook, Proc. of the IEEE 98, 2124-2154 (2010).

[29] T. Tahara, H. Koike, M. Kameno et al.: Room-Temperature Operation of Si Spin MOSFET with High On/Off Spin Signal Ratio, Appl. Phys. Express 8, 11304 (2015).

[30] V. Sverdlov, S. Selberherr: Silicon Spintronics: Progress and Challenges, Physics Reports, 585, 1-40 (2015).

[31] G. Schmidt, D. Ferrand, L.W. Molenkamp et al.: Fundamental Obstacle for Electrical Spin Injection from a Ferromagnetic Metal into a Diffusive Semiconductor, Phys. Rev. B 62, R4790-R4793 (2000).

[32] E.I. Rashba: Theory of Electrical Spin Injection: Tunnel Contacts as a Solution of the Conductivity Mismatch Problem, Phys. Rev. B 62, R16267-R16270 (2000).

[33] Y. Bychkov, E. Rashba: Properties of a 2D Electron Gas with Lifted Spectral Degeneracy, JETP Lett. 39, 78 (1984).

[34] P. Chuang, S.-C. Ho, L.W. Smith et al.: All-Electric All-Semiconductor Spin Field-effect Transistors, Nature Nanotechnology 10, 35 (2015).

[35] J. Ingla-Aynes, F. Herling, J. Fabian et al.: Electrical Control of Valley-Zeeman Spin-Orbit-Coupling–Induced Spin Precession at Room Temperature, Phys.Rev.Lett. 127, 047202 (2021).

[36] S. Ikeda, J. Hayakawa, Y. Ashizawa et al.: Tunneling Magnetoresistance of 604% at 300K by Suppression of Ta Diffusion in CoFeB/MgO/CoFeB Pseudo-Spin-Valves Annealed at High Temperature, Appl. Phys. Lett. 93, 082508 (2008).

[37] J .Slonczewski: Current-Driven Excitation of Magnetic Multilayers, J. Magn. Magn. Mater. 159, L1-L7 (1996).

[38] L. Berger: Emission of Spin Waves by a Magnetic Multilayer Traversed by a Current, Phys. Rev.B 54, 9353-9358 (1996).

[39] D. Apalkov, B. Dieny, and J.M. Slaughter: Magnetoresistive Random Access Memory, Proc. of the IEEE 104, 1796-1830 (2016).

[40] J. G. Alzate, U. Arslan, P. Bai, J. Brockman, Y. J. Chen et al.: 2Mb Array-Level Demonstration of STT-MRAM Process and Performance towards L4 Cache Applications, in Proc. IEDM Conf., 2019, pp.2.4.1–2.4.4.

[41] Y.-D. Chih, C.-C. Chou, Y.-C. Shih et al.: Design Challenges and Solutions of Emerging Nonvolatile Memory for Embedded Applications, in Proc. IEDM Conf., 2021, pp.2.4.1–2.4.4.

[42] H. Sato, M. Yamanouchi, S. Ikeda et al.: MgO/CoFeB/Ta/CoFeB/MgO Recording Structure in Magnetic Tunnel Junctions with Perpendicular Easy Axis, IEEE Transactions on Magnetics 49, 4437–4440 (2013).

[43] K. Nishioka, H. Honjo, S. Ikeda et al.: Novel Quad Interface MTJ Technology and its First Demonstration with High Thermal Stability and Switching Efficiency for STTMRAM beyond 2Xnm, in 2019 Symposium on VLSI Technology, 2019, pp.T120–T121.

[44] B. Jinnai, J. Igarashi, K. Watanabe et al.: High-Performance Shape-Anisotropy Magnetic Tunnel Junctions down to 2.3 nm, in 2020 IEEE International Electron Devices, in Proc. IEDM Conf., 2020, pp.24.6.1–24.6.4.

[45] S. Fiorentini, J. Ender, R. Orio et al.: Comprehensive Evaluation of Torques in Ultra Scaled MRAM Devices, in 2022 SISPAD, accepted.

[46] T. Hadamek, S. Fiorentini, M. Bendra et al.: Temperature Increase in STT-MRAM at Writing: A Fully Three-Dimensional Finite Element Approach; Solid-State Electronics, 193 108269-1—7 (2022).

[47] T. Hadamek, W. Goes, S. Selberherr et al.: Modeling Thermal Effects in STT-MRAM, in 2022 SISPAD, accepted.

[48] W. Loch, S. Fiorentini, N.P. Jorstad et al.: Double Reference Layer STT-MRAM Structures with Improved Performance, Solid-State Electronics 194, 108335-1—4 (2022).

[49] G. Hu, G. Lauer, J. Z. Sun at al.: 2x Reduction of STT-MRAM Switching Current Using Double Spin-Torque Magnetic Tunnel Junction, in Proc. IEDM Conf., 2021, pp.43-46.

[50] N.P. Jorstad, S. Fiorentini, W. Loch et al.: Finite Element Modeling of Spin-Orbit Torques, Solid-State Electronics 194, 108323-1—4 (2022).

A Novel Ferroelectric Nanopillar Multi-level Cell Memory

Hyeongu Lee and Mincheol Shin*
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
*mshin@kaist.ac.kr

Abstract

In this work, we present the novel multi-state non-volatile memory (NVM) device where ferroelectric (FE) nanopillars are embedded in a dielectric (DE) medium. Using our in-house 3D phase field simulator developed to treat the FE-DE composite system stably, we demonstrate that n FE nanopillars can generate more than 2^n states, enabling high storage capacity. The multi-states of the pillar array device are attributed to the depolarization field modulation with the pillar height and the multi-domain topological states of nanoscale FE structures.

Introduction

FE materials have attracted attention for NVM applications where two reversible polarization states are utilized [1]. To increase the storage capacity of the ferroelectric-based NVM, there have been reported various device concepts to achieve multi-states in a single cell. For example, the multi-layer structure of FE and DE can exhibit multi-states through sequential switching of each FE layer [2] or modulating the number of switching pathways [3]. These approaches generate multi-states that are fully poled FE states. In addition to using the poled states, multi-domain topological states of the nanoscale FE structures can be also utilized [4]. In this work, we present a novel multi-level NVM of FE nanopillars embedded in a dielectric medium. With 3D finite element TDGL simulations, we show that a 2-pillar system can create more than 4 states. Indeed, using a simple step pulse, we find that 8 states can be accessed from initial fully-poled polarization states.

Simulation methods

For a composite system of FE and DE, the polarization ($\boldsymbol{P}$) discontinuity occurs at the interface between FE and DE. The discontinuity results in the depolarization field which plays an important role in $\boldsymbol{P}$ switching. In addition, it can cause numerical instability with oscillations and divergence [5]. In this work, to accurately incorporate the depolarization field effect and resolve the issue of the numerical instability, we have developed 3D finite element phase-field simulator where TDGL equation is solved together with Poisson's equation. The time evolution of $\boldsymbol{P}$ is obtained by solving the TDGL equation as

$$-\rho\frac{dP_i(\boldsymbol{r},t)}{dt}=\frac{\delta F}{\delta P_i(\boldsymbol{r},t)}, \qquad i=x,y,z \quad (1)$$

$$F=\int_V dV \ (\alpha_{ij}P_iP_j+\beta_{ijkl}P_iP_jP_kP_l + \gamma_{ijklmn}P_iP_jP_kP_lP_mP_n+k_{ijkl}\partial_jP_i\partial_kP_l-P_iE_i) \quad (2)$$

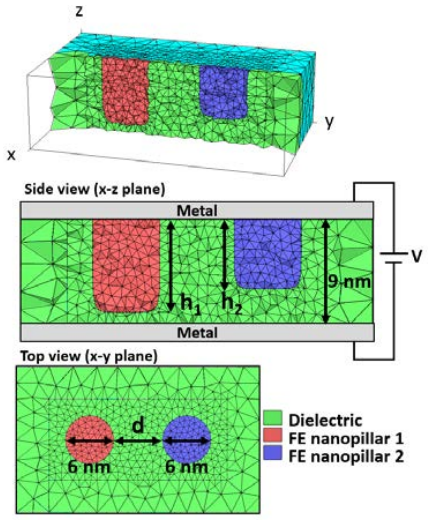


Fig. 1. Schematic structure of FE-nanopillar devices with the FEM mesh configuration

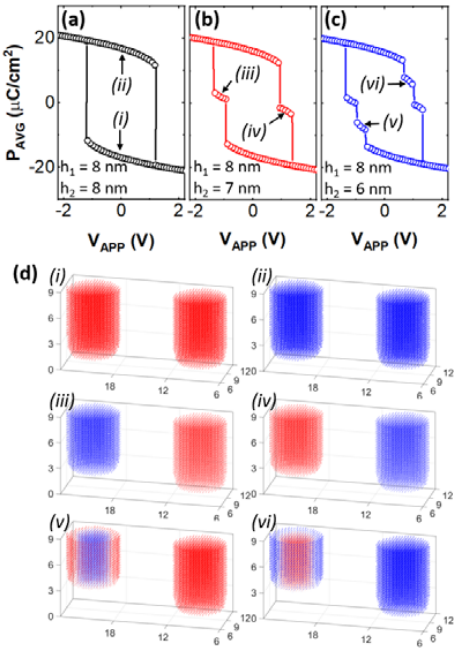


Fig. 2. P_{AVG} - V_{APP} hysteresis loops for two FE nanopillars in dielectric medium with different pillar heights (h_1 and h_2). (a) h_1= h_2= 8 nm, (b) h_1= 8 nm and h_2= 7 nm, and (c) h_1= 8 nm and h_2= 6 nm. (d) Polarization distribution of *(i)-(vi)* states of *(a)-(c)*. Blue and red colors indicates the up and down polarization, respectively.

where ρ is the kinetic coefficient, F is the total free energy, k_{ijkl} is the gradient energy coefficient, α_{ij}, β_{ijkl}, γ_{ijklm} are the Landau free energy coefficients, and E_i is the electric field. The in-house simulator can treat 3D geometrical model (x, y, z) and polarization vector (P_x, P_y, P_z). In this work, however, a uniaxial ferroelectric is assumed whose $\boldsymbol{P}$ direction

is parallel to z-axis. In general, the polarization values ($\boldsymbol{P}$) calculated from the TDGL equation are differentiated to obtain the polarization bound charges $\nabla \cdot \boldsymbol{P}$ for the Poisson's equation. Instead, to directly use $\boldsymbol{P}$ without a loss of precision by the numerical differentiation, the Poisson's equation is solved in mixed finite element formulation introducing the displacement vector ($\boldsymbol{D}$) as a unknown variable,

$$\nabla \cdot \boldsymbol{D}(\boldsymbol{r}) = 0 \quad (3)$$

$$\boldsymbol{D}(\boldsymbol{r}) - \epsilon(\mathbf{r})\nabla\psi(\boldsymbol{r}) - \boldsymbol{P}(\boldsymbol{r}) = \mathbf{0} \quad (4)$$

To reduce the number of unknown variables and avoid the numerical instability, the local discontinuous Galerkin finite element method is employed for the space discretization of Eq. (1), (3), and (4) [6]. The time evolution of Eq. (1) is calculated using the diagonally implicit Runge-Kutta method [7].

Results and Discussion

A single FE nanopillar can take at least two polarization configurations of the fully poled states. Thus, an array structure with n FE nanopillars can have 2^n states, as each nanopillar has two poled states. To investigate the multi-state property, we simulated the devices with two FE cylindrical nanopillars in the dielectric medium, where the heights, h_1 and h_2, of the pillar, are varied as shown in Fig. 1. The diameter of each pillar is 6 nm and inter-pillar distance (d) is kept fixed at 6 nm.

Figure 2 shows the simulated hysteresis loop of the average polarization of two pillars (P_{AVG}) as a function of the applied voltage (V_{APP}). In Fig. 2(a), only two P_{AVG} states are exhibited for the case of two identical pillars ($h_1 = h_2$). These states are the poled monodomain states shown in panels *(i)* and *(ii)* of Fig. 2(d). The polarization states of the two nanopillars simultaneously and independently switch because the coercive fields of the two pillars are same. If the pillar heights are different (h_1= 8 nm and h_2 = 7 nm), the two states labeled as *(iii)* and *(iv)* in Fig. 2(b) are generated besides the monodomain states. The two states result from the polarization switching of the shorter FE nanopillar as illustrated in panels *(iii)* and *(iv)* of Fig.2(d). This is because the depolarization field increases as the pillar height decreases, and the polarization state of the shorter pillar of h_2 = 7 nm is switched to the opposite poled state at weaker applied voltage.

It is remarkable that a further decrease in the pillar height of h_2 = 6 nm creates two additional states labeled as *(v)* and *(vi)* in Fig. 2(c). The states originate from the multi-domain topological states illustrated in panels *(v)* and *(vi)* of Fig. 2(d). These concentric-cylindrical states (CCS) are induced by the Landau-Lifshitz domain branching effect where the depolarization field from the nonuniform monodomain polarization distribution reduces the polarization at the center surface [5], [8].

To explore the possible hidden states which are not shown in the hysteresis loop, the step pulse is applied to the device with h_1= 8 nm and h_2 = 6 nm. Two initial states of up and down monodomain states are set up, and the pulse of the opposite electric field is applied. The pulse width is varied from $5\tau_k$ to $45\tau_k$, where τ_k is the relaxation time scale of $\rho/|\alpha|$ of Fig. 3(a) and (b). As shown in Fig. 3(c) and (d), the six stable states of *(i)-(iv)* correspond to those in Fig. 2(d). Note that two hidden states of *(vii)* and *(viii)* are accessed by the simple step pulse, and these states are created by the CCSs of the longer pillar of h_1= 8 nm, as shown in panels *(vii)* and *(viii)* of Fig. 3(e).

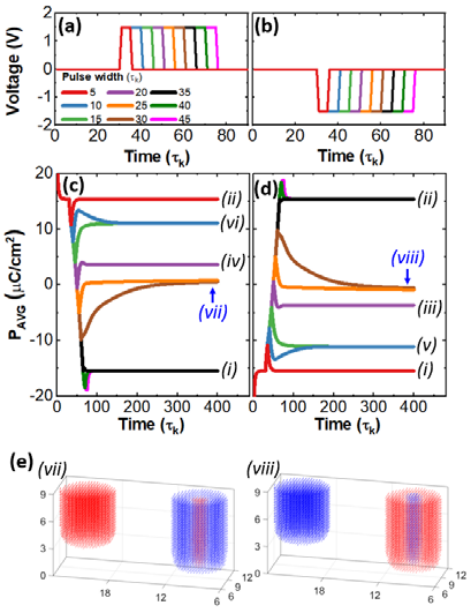


Fig. 3. Step pulse of (a) 1.5 V and (b) -1.5 V with different pulse widths. $\boldsymbol{P_{AVG}}$ as a function of pulse widths for the initial monodomain states of (c) up and (d) down. *(i)-(vi)* states correspond to those in Fig. 2(d). (e) The hidden states of *(vii)* and *(viii)*.

Conclusions

In this work, we propose a novel multi-level NVM using FE nanopillars in a dielectric medium. Through 3D TDGL simulations, we theoretically demonstrate that n-nanopillars can generate more than 2^n states. The multi-states of the FE pillar system are attributed to the consecutive switching of the polarization states of the FE nanopillars with different heights and the topological multi-domain states. This feature can be used toward high storage capacity NVM.

Acknowledgment

This work was supported by Samsung Electronics Co., Ltd. (IO201218-08230-01)

References

[1] M. Kim et al., IEEE Electron Device Lett., vol. 42, no. 11, 2021
[2] K. Ni et al., IEDM, pp. 28-9, 2019.
[3] R. Xu et al., Nat. Commun., vol. 10, no. 1., pp. 1-10, 2019
[4] P.-W. Martelli et al., Europhys. Lett., vol. 111, p. 50001, 2015
[5] P. Lenarczyk at al., SISPAD, pp. 311-314, 2016.
[6] P. Castillo et al., SIAM J. Numer., Anal., vol. 38, no. 5, pp. 1676-1706, 2000
[7] C. A. Kennedy et al., Appl. Numer. Math., vol. 146, pp. 221-244, 2019
[8] L. D. Landau and E. M. Lifshitz, Electrodynamics of continuous media (Elsevier, New York), 1985

Semi-empirical and VerilogA compatible compact model for ferroelectric hysteresis behavior

M. Lederer*, R. Olivo*, N. Yadav*, S. De*, K. Seidel*, L. M. Eng†, T. Kämpfe*
*Fraunhofer IPMS, CNT, Dresden 01099, Germany; Email: maximilian.lederer@ipms.fraunhofer.de
‡TU Dresden, Dresden 01099, Germany;

***Abstract*—This paper reports a semi-empirical, SPICE compatible and computationally efficient compact model for ferroelectric capacitors (Fe-CAP). This compact model is inspired by Jiles-Atherton model of ferromagnets, which features significantly smaller computational effort than other state-of-the-art models. This model successfully reproduces the evolution of the memory window and hysteresis of ferroelectric capacitors for any arbitrary signal. We have successfully calibrated this model with the experimentally characterized polarization switching dynamics of fabricated 10nm silicon doped hafnium oxide based Fe-CAP.**

***Index Terms*—Hafnium oxide, compact modeling, ferroelectric capacitors, nonvolatile memory, FeFET.**

I. Introduction

Due to its CMOS compatibility and scalability, hafnium oxide (HfO_2) based FeFET are been investigated as potential next generation nonvolatile memory (NVM). Successful integration of FeFETs into advanced technology nodes have further triggered the research interest among the scientists for deploying it with embedded NVM applications. Hereforth, it is necessary to have a SPICE compatible, computationally efficient compact model to investigate the performance of ferroelctric capacitors for large-scale integration. There have been numerous studies on compact model generation for Fe-CAPs. The very first physical compact model is based on the phenomenological Landau equation, which describes the ferroelectric behavior by a power-law relation of polarization and voltage. However, this relationship is not always physically achievable. Theoretically, this model predicts a negative capacitance region, which is suited for steep-slope FETs, however not appropriate for memory cells such as FeFET.Previously mainly applied FeFET memory models, which are based on the Preisach model are computationally demanding, which limit their applications in large circuit implementations, such as neuromorphic systems. Here, we present a computationally efficient ferroelectric model based on Jiles-Atherton equations, which is Verilog-A compatible and also reproduces the experimental results with high accuracy.

II. Experimental

For comparison a 10nm ferroelectric Si:HfO2 has been prepared. After the initial deposition of a TiN bottom electrode, the Si:HfO_2 layer is deposited using atomic layer deposition with a Hf:Si cycling ratio of 16:1. After the deposition of a TiN top electrode, the film is crystallized using rapid thermal annealing. Capacitors are formed by Ti/Pt shadow mask deposition and subsequent wet etch. Dynamic hysteresis measurements are performed at 1 kHz using an Aixacct TF3000 analyzer.

This research was funded by the ECSEL Joint Undertaking project TEMPO in collaboration with the European Union's Horizon 2020 Framework Program for Research and Innovation (H2020/2014-2020) and National Authorities, under Grant No. 826655.

The ferroelectric compact model is coded in Python 3 as well as Verilog-A and simulation has been done in Spectre 6.1.8 circuit simulator using the Verilog-A code. It is a ferroelectric capacitor model only, hence it can be included onto a BSIM model. The ferroelectric thickness, permittivity, and fitting parameter can be declared within the model file or these can be included during the simulation. It has two excess ports (in and out) for the circuit simulation. For circuit simulation, a ferroelectric capacitor (symbol) is called in the schematic editor then a supply voltage and one Ohm resistance are connected in series. The change in current at the entry node is measured with supply voltage. The transient simulation results are collected.

III. Modelling of a ferroelectric hysteresis

For deeper understanding of a ferroelectric device, it is essential to be able to model it in all possible conditions of operation. For such a model, a semi-empirical model based on physical equations that can correlate material properties with voltage-current characteristics, while giving expected results, was implemented. Ferroelectric behavior is usually solved by calculating the polarization (charge), from which one has to use the derivative to obtain the current characteristics of the device.

Diverse models attempt to describe this behavior, but are showing disadvantages: Hyperbolic tangent based models are only compatible with full-loop operation [5]; Preisach density based models require a large set recurring time-series parameters [4], [7], [9] and finally, models based on nucleation limited switching [6] have the same problem and are more complex to be extended to arbitrary signals [1], [3], [8].

For this reasons, a model developed to described ferromagnetic films based on the $Langevine$ function, commonly referred to as Jiles-Atherton model (based on the original authors) [2] was chosen. Such a model is compatible with circuit simulators, supports various operation points and can be adapted to different devices, since it's based on semi-empirical parameters. Nonetheless, a ferroelectric analogue has not been proposed so far. By implementation of the displacement field

D (see equ. 1), contributing polarization P_0 from the material, the permittivity of free space ϵ_0, the applied electric field E and the Langevine equation (see equ. 2), which relates the Brillouin equation to an countless amount of domains, one can create a solvable equation.

$$D = P_0 + \epsilon_0 E \tag{1}$$

$$L(x) = coth(x) - \frac{1}{x} \tag{2}$$

The polarization has to account for the hysteretic (or irreversible) and anhysteretic behavior, as such, the flow of the solution is splitted in two. The Langevine equation is used to solve the anhysteretic polarization P_{anh} (see equ. 3), while the irreversible polarization P_{irr} is solved by accounting for the whole polarization P and subtracting the anhysteretic part (see equ. 4). The total polarization inside the device is the sum of the ferroelectric and dielectric contributions based on the relative permittivity.

$$P_{anh} = P_S L\left(\frac{D_{eff}}{\gamma}\right) \tag{3}$$

$$P = (1-c)P_{irr} + cP_{anh} \tag{4}$$

The proposed model uses semi-empirical values obtained by fitting measured devices to describe physical qualities, as shown in Fig. 1 for the herein described model. Such a fitting approach allows for flexibility depending on the measured results. Conventional parameters such as the dielectric constant and thickness of the layer are set directly. The parameter that accounts for the amount of polarization due to ferroelectric domains compared to the internal polarization is denoted c. The saturation polarization P_S is the polarization at which the polarization-voltage curve would saturate. The parameter k affects the mean coercive field, γ its distribution and α the inter-domain coupling.

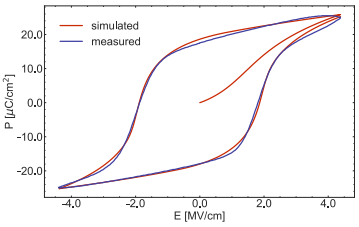


Fig. 1. Comparison of the polarization response of the simulated and measured ferroelectric capacitor. The simulation parameters were optimized via fitting to the experimental data. Minor differences are apparent due to the presence of leakage current in the measured hysteresis.

A. Implementation in VerilogA

As shown in Fig. 2a, the simulation using Verilog-A in Spectre converges numerically stable and the displacement current and displacement current peaks are clearly observable. Furthermore, Fig. 2b visualizes the subloop functionality of

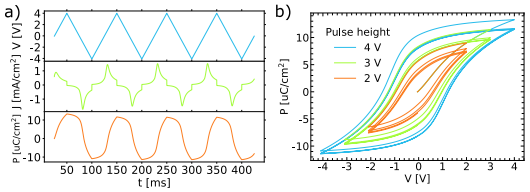


Fig. 2. Verilog-A based simulation of a ferroelectric capacitor. The current and polarization response over time is shown for multiple triangular voltage pulses in (a). The sub-loop behavior is illustrated in (b) for different voltage amplitudes.

the model. Consequently, this model can be used with arbitrary signals and the ferroelectric capacitor compact model can be introduced into various circuit simulations.

B. MFM and MFIS capacitors

One important application, where a ferroelectric compact model is required, is the ferroelectric field effect transistor (FeFET). The major difference to the above simulated metal-ferroelectric-metal (MFM) capacitor is that the FeFET gate stack is described by a metal-ferroelectric-insulator-semiconductor (MFIS) stack. Such device can be modeled as a ferroelectric capacitor on top of a regular transistor. This approach decouples the phenomena in the device, allowing the use of a capacitor model with a preexisting transistor, e.g. a proprietary design. A simple MFIS capacitor can be described analytically by equ. 5 based on the potential across each layer.

$$V_{ext} = \Phi_{MS} + \Phi_S(V_{FE}) + V_{IL}(V_{FE}) + V_{FE} \tag{5}$$

IV. Conclusion

In conclusion, it was demonstrated that the ferromagnetic Jiles-Atherton model can be transferred to a ferroelectric model. This semi-empirical model allows for arbitrary signals, enables sub-loop operation and does not require extensive history parameters. Furthermore, the integration of this model into Verilog-A and a Spectre circuit simulation environment has been demonstrated. Finally, it has been explained how this model can be applied for the simulation of FeFETs and the analytical integration into a MFIS capacitor model has been proposed.

References

[1] C. Alessandri et al., *IEEE Transactions on Electron Devices*, 1–8, 2019.
[2] D. C. Jiles et al., *Journal of Magnetism and Magnetic Materials 61*, 48–60, 1986.
[3] Y. Li et al., *Applied Physics Letters 114*, 142902, 2019.
[4] I. D. Mayergoyz et al., *IEEE Transactions on Magnetics 24*, 212–217, 1988.
[5] S. L. Miller et al., *Journal of Applied Physics 72*, 5999–6010, 1992.
[6] H. Mulaosmanovic, et al., *ACS Applied Matererials & Interfaces 9*, 3792–3798, 2017.
[7] K. Ni et al., *IEEE Symposium on VLSI Technology*, 2018, 131–132.
[8] S. Oh et al., *APL Materials 7*, 091109, 2019.
[9] P. Wang et al., *IEEE Transactions on Electron Devices 67*, 1–6, 2020.

New Insights into the Effect of Spatially Distributed Polarization in Ferroelectric FET on Content Addressable Memory Operation for Machine Learning Applications

Chang Su[1], Weikai Xu[1], Qianqian Huang[1,2*], Lining Zhang[3] and Ru Huang[1,2*]
[1]School of Integrated Circuits, Peking University, Beijing 100871, China.
[2]National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Beijing 100871, China.
[3]School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China.
([*]Email: hqq@pku.edu.cn; ruhuang@pku.edu.cn)

***Abstract*— In this work, the impacts of spatially distributed polarization in Ferroelectric FET (FeFET) on the performance of content addressable memory (CAM) circuits are investigated. It is found that for CAM operation with large pre-charging voltage in the match line, the storage state in FeFET will be changed with the lower polarization near drain side, resulting in the strong non-uniformity of polarization along the channel direction. The lateral electric field generated by source and drain junction due to short channel effect and the device parameters variation due to multi-domain features will further exacerbate this effect. It is further shown that this distributed polarization effect may lead to significant circuit performance difference (search delay increased by 53.2% and search energy reduced by 53.6%) compared with uniform polarization situation, and may further degrade the search accuracy especially for multi-bit CAM. This work indicates the crucial role of polarization distribution effect in the device-circuit co-optimization of FeFET for the first time.**

Keywords*— *TCAD simulation, FeFET, CAM

I. INTRODUCTION

Content addressable memory (CAM) has drawn significant interests for data-centric machine learning (ML) applications due to its high parallel pattern matching capability [1-2], while conventional SRAM-based CAM suffers from large area overhead [3]. Various emerging non-volatile memories (NVM) has been applied to implement more compact cells [4-6]. Among them, Ferroelectric FET (FeFET)-based Ternary CAM (TCAM) with ultra-compact structure (2FeFET) [6] (Fig. 1a) significantly boosts the CAM density with the additional advantages of large on-off ratio and low write energy [2]. Moreover, the multi-level states caused by partial polarization switching of the ferroelectric (FE) layer enables the realization of FeFET-based multi-bit CAM (MCAM) design [7-9]. To evaluate the CAM circuit performance, the FeFET device model, established by directly solving the FE model with baseline MOSFET model, are usually utilized [10], which are intrinsically a lumped treatment for FeFET with internal metal layer structure [11]. However, for FeFET without the internal metal layer in the gate stack, the polarization (P_{FE}) and the FE voltage (V_{FE}) may vary spatially in the longitudinal direction at a nonzero drain bias (Fig. 1b) and can also be influenced by lateral field at drain/source edges [12]. Therefore, the polarization distributed along the channel direction in FeFET without internal metal layer is non-uniform, and may directly impact the device channel conduction and the FeFET-based CAM circuit performance.

In this work, for the first time, the impact of distributed polarization in FeFET device on the CAM circuit performance is investigated by simulation. Different from the FeFET with uniform polarization, the device with polarization distribution effect shows the significantly different conduction capability considering the high drain bias especially in the multi-bit CAM applications, indicating its crucial role in the device-circuit co-optimization of FeFET.

II. SIMULATION METHOD

To investigate the impact of distributed polarization on the CAM cell, the mixed mode simulation in Sentaurus TCAD [13] is carried out. For the FeFET simulation, the FE parameters are calibrated with [14] (Fig. 2a) and the polarization switching behavior is described by the Preisach model [15]. For CAM cell simulation, the different non-volatile polarization states are programmed to FeFET by the same-width pulses with different amplitudes, as shown in the device I_D-V_G curves with different threshold voltage (V_{TH}) values (Fig. 2). Then the match line (ML) is pre-charged to the voltage equal to VDD. In the searching phase, the relatively small gate voltage (V_G) with the related amplitudes (V_{SL}) are subsequently applied according to the search query. The truth table and operation results of FeFET-based CAM cell are dependent on the ML discharging speed during the search phase, which is directly decided by the conduction current (I_{ML}) of the parallel FeFET devices (Fig. 2&3).

III. RESULTS AND DISCUSSION

Based on the above discussion, the polarization state in FeFET device directly determines the search results of CAM cell. To investigate the impact of polarization distribution effect, FeFET with an internal metal layer is also simulated, showing the uniform P_{FE} due to the equipotential surface even with relatively short gate length (Fig. 4a). For FeFET without internal metal layer, due to the participation of lateral electric field generated by source and drain junction with the vertical field, both the electric potential and polarization are spatially distributed after program operation (Fig. 4b). Moreover, considering the CAM operation, when a high drain bias is applied during the pre-charging phase and will be sustained for match situation, it can be seen from Fig. 4c that the non-uniformity of electric field becomes stronger, leading to a significant P_{FE} change along the channel direction. Besides, as shown in the extracted polarization distribution in Fig. 5a, the lower P_{FE} is observed near the drain region than the center region, which is due to the stronger vertical field from junction [12] and the high drain bias. An intuitive representation of the coupling effect between electric potential and P_{FE} is shown in the load line plot in Fig. 5b. Therefore, the difference of transfer characteristics of FeFET (Fig. 6) will influence the discharging process of CAM mismatch (Fig. 7). In addition, it is shown that for conventional memory applications of FeFET with low V_{DS}, the impact of distributed P_{FE} is relatively small, while for the CAM applications, re-evaluation by device-circuit co-optimization considering the polarization distribution is necessary.

Moreover, the P_{FE} distribution impact will become more significant for MCAM application. It is shown that for intermediate state with partially switched P_{FE}, with the same V_{SL}, the channel current difference can be as much as several orders of magnitude due to the P_{FE} distribution effect (Fig. 8a), which will strongly degrade the search delay (Fig. 8) and thereby reduce the search accuracy within a limited search time (Fig. 9). Besides lateral field and drain bias, for practical FeFET, the P_{FE} distribution effect can also originate from the coexistence of multi-domain and multi-phase in FE layer with spatially distributed FE parameters and polarization direction [16-18], inducing the large V_{TH} variation of I_D-V_G curves and reducing the search accuracy of FeFET-based CAM (Fig. 10).

In order to evaluate the impacts of P_{FE} distribution on the circuit performance, a 64x64 FeFET CAM array is simulated and compared by HSPICE. As shown in Fig. 11, the search delay is increased by 53.2% when considering the P_{FE} distribution and thereby the related search energy is reduced by 53.6%, indicating the significant P_{FE} distribution impact for the design technology co-optimization of FeFET.

IV. CONCLUSION

This work simulates and investigates the impacts of spatially distributed polarization on the performance of FeFET-based CAM. Compared with the uniform polarization situation, the P_{FE} distribution may induce the larger V_{TH} and smaller current under a high drain bias in FeFET. For CAM operations with relatively larger drain bias than the conventional memory applications, the P_{FE} distribution will lead to significant difference in circuit performance, showing its crucial role in the device-circuit co-optimization.

ACKNOWLEDGEMENT

This work was supported by National Key R&D Program of China (2018YFB2202801), NSFC (61927901，61851401, 61822401), Beijing Nova Program of Science and Technology (Z191100001119101), BJSAMT Project (SAMT-BD-KT-22030101) and 111 Project (B18001).

References: [1] C. E. Graves, et al., AM, pp. 32, 2020. [2] K. Ni, et al., Nat. Electron., pp. 521-529, 2019. [3] K. Pagiamtzis, et al., IEEE JSSC, pp. 712–727, 2006. [4] J. Li, et al., IEEE JSSC, pp. 896-907, 2013. [5] C. Li, et al., Nat. Commun., pp. 1-8, 2020. [6] X. Yin, et al., IEEE TCAS-II, pp. 1577-1581, 2019. [7] X. Yin, et al., IEEE TED, pp. 2785-2792, 2020. [8] A. Kazemi, et al., DATE, pp. 1084-1089, 2021. [9] C. Li, et al., IEDM, pp. 29.3.1-29.3.4, 2020. [10] K. Ni, et al., VLSI, pp. 131-132, 2018. [11] J. P. Duarte, et al., IEDM, pp. 754-757, 2016. [12] S. Jindal, et al., IEEE TED, pp. 1364-1368, 2021. [13] Sentaurus Device User Guide Version: O-2018.06, Synopsys, Mountain View, CA, USA, June. 2018. [14] J. Muller, et al, NVMTS, pp. 1-7, 2016. [15] B. Jiang, et al., VLSI, pp. 141-142, 1997. [16] Y. S. Liu, et al., IEEE EDL, pp. 369-372, 2020. [17] K. Ni, et al., IRPS, pp.1-5, 2020. [18] G. Choe, et al., IEEE JEDS, pp.1131-1136, 2021.

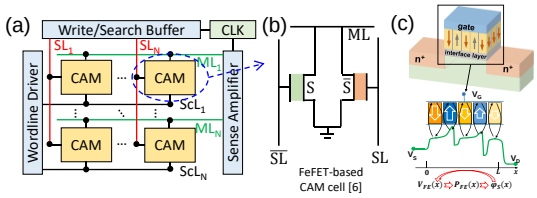


Fig. 1 (a) Architecture of the CAM array. (b) The basic structure of 2FeFET-based CAM cell [6]. (3) The 3D schematic structure of FeFET with spatially distributed polarization which shows local coupling with the underlying semiconductor.

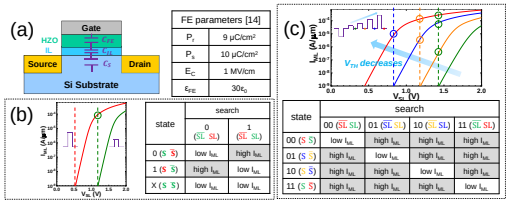


Fig. 2 (a) Schematic structure of the FeFET simulated in this work with FE parameters referred to [14]. Simulated I_D-V_G curves for (b) TCAM and (c) MCAM. The operation principles are shown in the truth table.

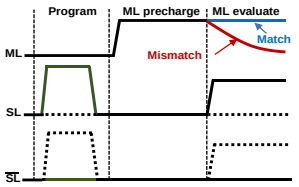


Fig. 3 The waveform of operation voltage in the CAM application.

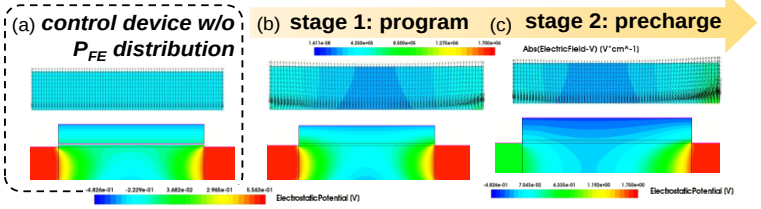


Fig. 4 Simulated results of the electric field distribution in the FE layer and electric potential along the channel for (a) the control device with uniform P_{FE} and the device with distributed P_{FE} after (b) program and (c) pre-charge operation.

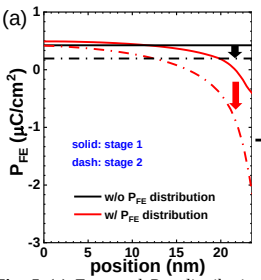


Fig. 5 (a) Extracted P_{FE} distribution near the drain junction. (b) The schematic load line plot between the local FE layer and underlying semiconductor respectively near the source and drain region, showing the different polarization switching behavior caused by nonzero drain bias.

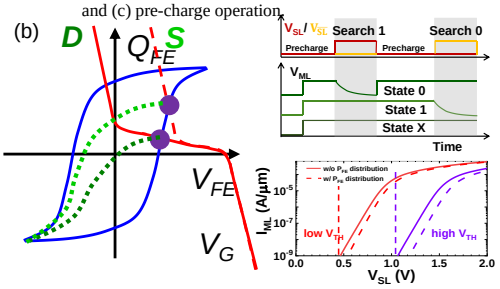


Fig. 6 The search operation of TCAM and the impacted I_D-V_G curves due to the distributed P_{FE}.

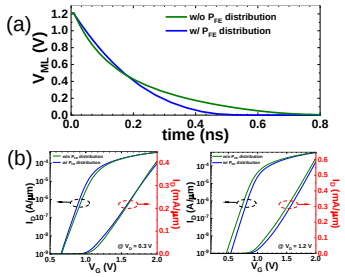


Fig. 7 (a) Different discharging behavior due to (b) the different channel conduction caused by the P_{FE} distribution.

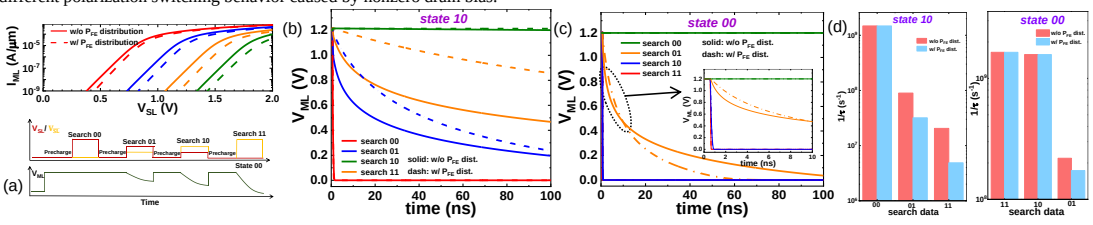


Fig. 8 (a) The search operation of 2-bit MCAM and a set of I_D-V_G curves showing the impact from distributed P_{FE}. Simulated search operation of MCAM when storing (b) state 10 and (c) state 00. (d) The discharging speed respectively corresponding to state 10 and state 00.

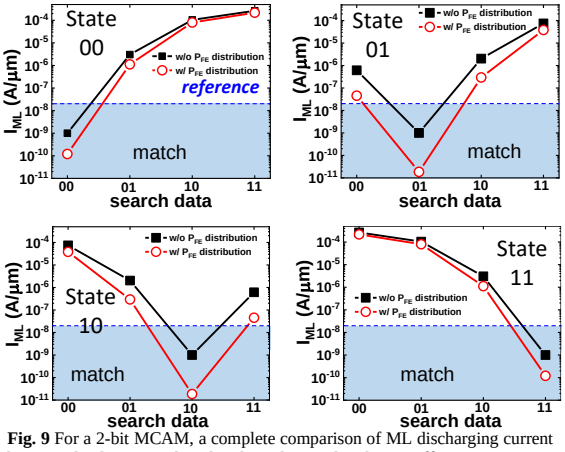


Fig. 9 For a 2-bit MCAM, a complete comparison of ML discharging current between the devices with and without the P_{FE} distribution effect.

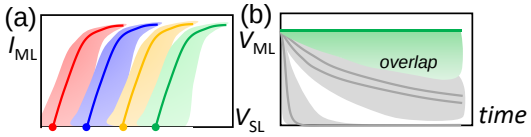


Fig. 10 The schematic diagram of the impact from variation on the (a) I_D-V_G curves and (b) search operation for MCAM.

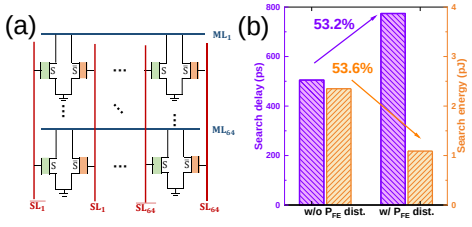


Fig. 11 (a) The structure of the simulated 64x64 array. (b) Performance evaluation considering the effect of distributed polarization.

Comprehensive Evaluation of Torques in Ultra Scaled MRAM Devices

Simone Fiorentini[1,2], Johannes Ender[1,2], Roberto L. de Orio[2],
Siegfried Selberherr[2], Wolfgang Goes[3], and Viktor Sverdlov[1,2]
[1] Christian Doppler Laboratory for Nonvolatile Magnetoresistive Memory and Logic
at the [2]Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, A-1040 Wien, Austria
[3] Silvaco Europe Ltd., Cambridge, United Kingdom

Abstract—**We present a generalization of the coupled spin-charge drift-diffusion formalism capable of accurately describing the spin and charge transport properties through magnetic tunnel junctions. Additional correction terms enable reproducing oscillations of the spin current in ferromagnets typical for quasi-ballistic transport. Our approach proves necessary to accurately capture an interplay between the interfacial Slonczewski and bulk-like Zhang-Li contributions to the torque in ultra-scaled MRAM devices.**

Keywords – Spin and charge drift-diffusion, spin-transfer torque, STT-MRAM

I. Introduction

In recently demonstrated ultra-scaled MRAM devices, elongated composite free layers with multiple MgO barriers are employed to boost both the shape- and interface-induced perpendicular magnetic anisotropy [1]. Therefore, when simulating ultra-scaled STT-MRAM, it is paramount to describe both the Slonczewski [2] and Zhang-Li [3] contributions to the torque. This requirement is achieved by employing the spin drift-diffusion approach for the computation of the spin accumulation $\mathbf{S}$ in the free layer of the structure [4], [5] with the equations

$$\mathbf{J_S} = -\frac{\mu_B}{e}\beta_\sigma \mathbf{m} \otimes \left(\mathbf{J_C} - \beta_D D_e \frac{e}{\mu_B}[(\boldsymbol{\nabla}\mathbf{S})\mathbf{m}]\right) - D_S \boldsymbol{\nabla}\mathbf{S}, \quad (1a)$$

$$\frac{\partial \mathbf{S}}{\partial t} = -\boldsymbol{\nabla}\mathbf{J_S} - D_S\frac{\mathbf{S}}{\lambda_{sf}^2} - \mathbf{T_S} = \mathbf{0}, \text{and} \quad (1b)$$

$$\mathbf{T_S} = -\frac{D_e}{\lambda_J^2}\mathbf{m}\times\mathbf{S} - \frac{D_e}{\lambda_\varphi^2}\mathbf{m}\times(\mathbf{m}\times\mathbf{S}), \quad (1c)$$

where μ_B is the Bohr magneton, e is the electron charge, β_σ and β_D are polarization parameters, D_e is the electron diffusion coefficient, λ_{sf} is the spin-flip length, $\lambda_J$is the exchange length, λ_φ is the spin dephasing length, $\mathbf{J_C}$ is the charge current density, $\mathbf{J_S}$ is the spin current density tensor, $\mathbf{T_S}$ is the spin torque, and $\mathbf{m}$ is the unit magnetization vector. This formalism must be extended to accurately describe the torques acting in a magnetic tunnel junction (MTJ).

II. Model

The charge current through the tunnel barrier (TB) is modeled by employing a low conductivity which locally depends on the relative magnetization vector's orientation across the tunnel layer [6]. For the spin current, the diffusion coefficient in the TB is set low, proportionally to the conductivity, and the following boundary condition is imposed on the left and right interfaces of the TB:

$$\mathbf{J_S}\cdot\mathbf{n} = -\frac{\mu_B}{e}\frac{\mathbf{J_C}\cdot\mathbf{n}}{1+P_{RL}P_{FL}\mathbf{m_{RL}}\cdot\mathbf{m_{FL}}}\Big(\alpha P_{RL}\mathbf{m_{RL}} + \alpha P_{FL}\mathbf{m_{FL}} + \frac{1}{2}\left(P_{RL}P_{RL}^{\eta} - P_{FL}P_{FL}^{\eta}\right)\mathbf{m_{RL}}\times\mathbf{m_{FL}}\Big) \quad (2)$$

Here, the subscript RL(FL) indicates the reference (free) layer, P is the Slonczewski polarization parameter [2], P_{RL}^{η} is the out-of-plane polarization parameter [7], and α describes the influence of the interface spin-mixing conductance on the transmitted in-plane spin current [8]. Thereby, we can reproduce the spin current value [7] when $\mathbf{J_C}$ flows through the TB.

III. Results

While employing (2) gives the opportunity to fix the spin current density in the TB to the value expected in MTJs, the length parameters entering (1) determine the scale of absorption of the transverse spin accumulation components and the behavior of the torque in the bulk of the ferromagnetic (FM) layers. Fig. 1a reports the spin accumulation obtained in a symmetrical MTJ structure with $\lambda_J = 1$ nm, $\lambda_\varphi = 2$ nm, and $\lambda_{sf} = 10$ nm, where the FM layers are 2 nm thick. The magnetization points towards **x** in the reference layer (RL) and towards **z** in the free layer (FL). In this case, the transverse spin accumulation components are not completely absorbed in the FL, contrary to what is usually expected in strong ferromagnets [2], [9]. By taking an effective dephasing length of $\lambda_\varphi = 0.4$ nm, it is possible to have a faster decay of the transverse components close to the TB interface (cf. Fig. 1b). The sinusoidal angular dependence of the STT torque on the angle between the magnetization vectors in the FM layers predicted in MTJs under a constant voltage is reproduced exactly, as shown in Fig. 2. By introducing a ballistic correction [10] to the spin drift-diffusion formulation, a more complex torque behavior with oscillations is obtained. Fig. 3 shows the torque computed using $\lambda_J = 1$ nm, $\lambda_\varphi = 4.3$ nm, and the mean free path $\lambda = 5.8$ nm for semi-infinite FM layers, which qualitatively agrees with ballistic results [7].

Domain walls or magnetization textures are formed during switching in elongated FLs. In this case, both Slonczewski and Zhang-Li contributions are present. In Fig. 4 we show the torques acting in a 15 nm long FL with a textured magnetization, compared to the ones computed using the Zhang-Li expression modified to include the spin dephasing length λ_φ. The comparison reveals a substantial difference. The reason for the discrepancy lies in the fact that the presence of the TB also generates a weakly decaying spin accumulation component parallel to the magnetization, whose interaction with the magnetization texture substantially modifies the Zhang-Li contribution.

IV. Conclusion

We presented an extension of the drift-diffusion formalism capable of reproducing expected properties of the torque in magnetic tunnel junctions. Our modeling approach clearly demonstrates that in the presence of an MTJ the Slonczewski and Zhang-Li contributions are not independent, and that a unified treatment of the torque is needed in order to accurately describe the switching process in ultra-scaled MRAM cells presenting elongated and composite ferromagnetic layers.

Acknowledgment

The financial support by the Austrian Federal Ministry for Digital and Economic Affair, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

References

[1] B. Jinnai, J. Igarashi, K. Watanabe, T. Funatsu, H. Sato *et al.*, "Highperformance shape-anisotropy magnetic tunnel junctions down to 2.3 nm," in *Proc. IEDM Conf.*, pp. 24.6.1–24.6.4, 2020.

[2] J. C. Slonczewski, "Currents, torques, and polarization factors inmagnetic tunnel junctions," *Phys. Rev. B*, vol. 71, p. 024411, 2005.

[3] S. Zhang and Z. Li, "Roles of nonequilibrium conduction electrons on the magnetization dynamics of ferromagnets," *Phys. Rev. Lett.*, vol. 93, p. 127204, 2004.

[4] C. Abert, M. Ruggeri, F. Bruckner, C. Vogler, G. Hrkac *et al.*, "A three-dimensional spin-diffusion model for micromagnetics," *Sci. Rep.*, vol. 5, p. 14855, 2015..

[5] S. Lepadatu, "Unified treatment of spin torques using a coupled magnetisation dynamics and three-dimensional spin current solver," *Sci. Rep.*, vol. 7, p. 12937, 2017.

[6] S. Fiorentini, J. Ender, S. Selberherr, R. L. de Orio, W. Goes, and V. Sverdlov, "Coupled spin and charge drift-diffusion approach applied to magnetic tunnel junctions," *Sol.-St. El.*, vol. 186, p. 108103, 2021.

[7] M. Chshiev, A. Manchon, A. Kalitsov, N. Ryzhanova, A. Vedyayev *et al.*, "Analytical description of ballistic spin currents and torques in magnetic tunnel junctions," *Phys. Rev. B*, vol. 92, p. 104422, 2015.

[8] K. Y. Camsari, S. Ganguly, D. Datta, and S. Datta, "Physics-based factorization of magnetic tunnel junctions for modeling and circuit simulation," in *Proc. IEDM Conf.*, pp. 35.6.1–35.6.4, 2014.

[9] A. Brataas, G. E. W. Bauer and P. J. Kelly, "Non-collinear magnetoelectronics," *Phys. Rep.*, vol. 427, pp. 157-255, 2006.

[10] P. Graczyk and M. Krawczyk, "Nonresonant amplification of spin waves through interface magnetoelectric effect and spin-transfer torque," *Sci. Rep.*, vol. 11, p. 15692, 2021.

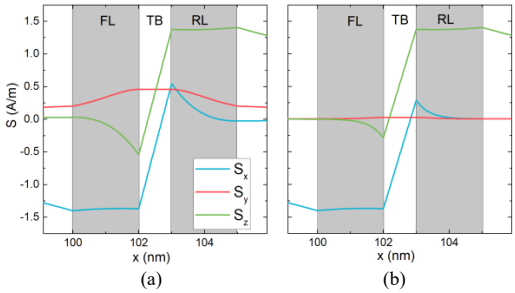


Fig. 1. Spin accumulation in symmetric MTJ structure. Nonmagnetic contacts are present to let **S** decay. The presence of the TB crates a jump in the **S** components. (a) Results for $\lambda_\varphi = 2$ nm. (b) Results for $\lambda_\varphi = 0.4$ nm.

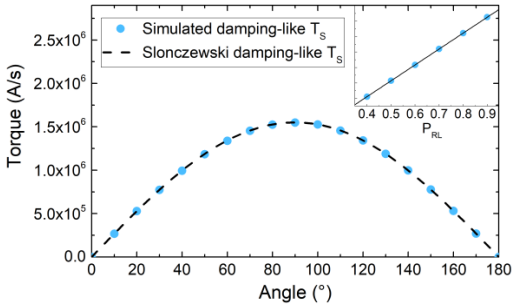


Fig. 2. Dependence of the average torque on the relative angle between the magnetization vectors. The inset shows the linear dependence of the torque on the RL polarization factor.

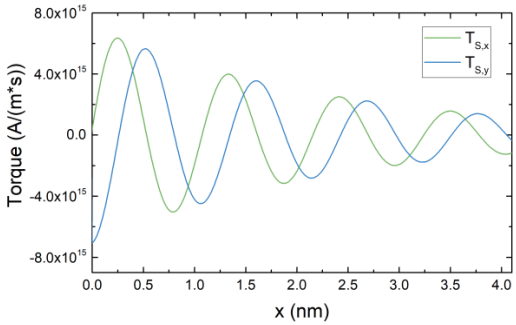


Fig. 3. Torque computed with the inclusion of ballisitic corrections to the spin current in a semi-infinite FL.

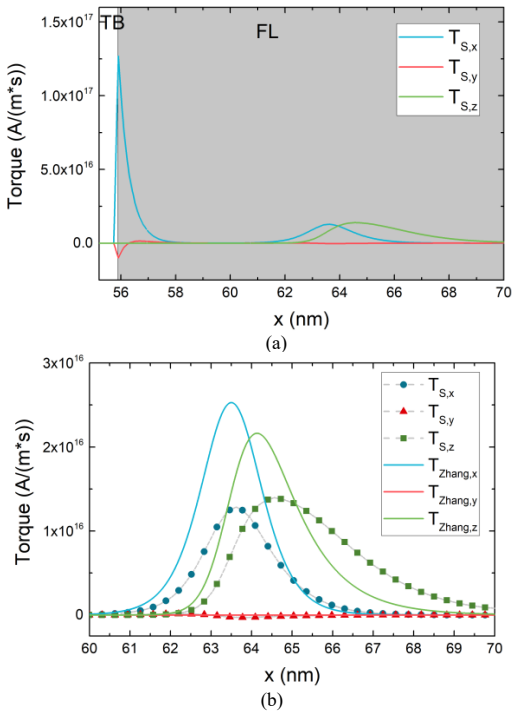


Fig. 4. (a) Spin torque in an elongated FL with the magnetization going from z to -x. The magnetization in the RL is along x. (b) Comparison of the spin torque to the Zhang-Li expression.

Analysis of Uniaxial Stress Impact on Drift Velocity of 4H-SiC by Full Band Monte Carlo Simulation

T. Nishimura, K. Eikyu, K. Sonoda and T. Ogata
Renesas Electronics Corporation, 751, Horiguchi, Hitachinaka, Ibaraki, 312-8511, Japan
Email: tomoya.nishimura.uf@renesas.com

***Abstract* — The stress response of the drift velocity of 4H-SiC is analyzed by Full band Monte Carlo simulation. The response decreases with an increase in the electric field except a hump around 1 MV/cm. The decreasing trend is explained by increasing scattering rates which diminish the effect of the change of effective mass due to stress. The hump comes from the transition of electrons to band minima with a lighter effective mass.**

I. Introduction

Si power devices have nearly reached the performance limit due to the physical characteristics of Si. Therefore, in recent years, wide bandgap materials, such as SiC, have been attracting attention.

To accurately reproduce the electron transport in a power device, the conventional simulation method considering only the band bottom is not sufficient. It is because under a high electric field, electrons transit to a higher-order band. So Full band Monte Carlo (FBMC) simulation is effective. There are many reports of analyzing the electron transport in SiC by FBMC simulation [1], but few have considered the effects of stress. Therefore, we have recently reported the FBMC analysis of the stress response of 4H-SiC in the c-axis direction [2] where the stress dependence of the impact ionization coefficient is small and the drift velocity is positively correlated with the tensile stress. In this paper, we analyze the stress response of the drift velocity in more detail.

II. Method

The band structure of 4H-SiC (Fig. 1) is calculated by the first-principal calculation tool DMol3 [3] and calibrated to correct the bandgap and band shape. The uniaxial stress is applied in the c-axis direction of the crystal, and lattice constants are set from the Young's modulus and Poisson's ratio. The FBMC simulator considers acoustic phonon scattering, optical phonon scattering, and impact ionization as scattering factors. And this simulator adopts double Brillouin zones for the calculation k-space domain which enables to use an orthogonal k-space mesh (Fig. 1). The details of the simulation parameters and calibration are reported elsewhere [2] [4].

III. Results and Discussion

Figures 2 and 3 show the c axis stress dependence of the band structure of 4H SiC by first principles calculations. The curvature of the conduction band bottom (M-point) decreases under compressive stress and increase under tensile stress. On the other hand, the bandgap decreases under tensile stress.

Next, the electric field dependence of the electron drift velocity is shown in Fig. 4. The electric field direction is the same as the stress application direction of the c-axis. At low electric fields, the drift velocity is proportional to the electric field, but after peaking once, it decreases at high electric fields. This can be explained by the band population. At low electric fields, the electrons are localized at the M- point of the conduction band bottom as shown in Fig. 5, and at high electric fields, they transit to other valleys, where electrons have a heavier effective mass than M-point. Therefore, the stress dependence of the drift velocity at a low electric field is mostly corresponding to the curvature change at M-point (Fig. 7).

Next, stress sensitivity coefficient of drift velocity γ is plotted against applied electric field in Fig. 8. The sensitivity between 10 GPa and 0 GPa is same tendency that of 0 GPa and -10 GPa. So, we discuss the sensitivity between 10 GPa and 0 GPa and defined as in the formula below with drift velocity V_d and applied stress σ.

$$\gamma = \frac{\Delta V_d \,/\, V_d@0\mathrm{GPa}}{\Delta\sigma}$$

In the low electric field region (i), the electrons are localized at M-point, so that γ is mostly corresponding to the effective mass change. Furthermore, in the region (ii) where the electric field becomes higher, the influence of the scatterings become larger, so that γ becomes smaller. An interesting feature is the peak appearing in the region (iii) around 1 MV/cm. This indicates that the slope becomes gentler under tensile stress. The electron population change in each band in this region is shown in Fig. 10 for analysis. Under no stress applied, the electrons localized near the M valley transit to the L-point or higher-order band with higher energy as the electric field increases. On the other hand, when tensile stress is applied, there are fewer electrons that transit to the L-point and higher-order bands, and more electrons transit to the H-point instead. So, the decrease in the drift velocity is alleviated since the effective mass of the electron located at the H-point is smaller than for the one at the L-point or the higher band. Furthermore, the cause of such a difference in band transition is a change in the band structure when stress is applied. Under tensile stress, the energy difference between points M and L at the conduction band bottom, and that between points M in the nearest higher band are wider. On the other hand, the energy difference from M-point to H-point shrinks. Therefore, it is considered that electrons are likely to transit to the H-point under tensile stress.

Furthermore, in the region (iv) where the electric field is large, impact ionization becomes apparent, but its influence on γ is small.

IV. Conclusion

The drift velocity and its stress response in low electric fields strongly depend on the curvature at M-point, but in high electric fields the response becomes small due to the scatterings. However around 1 MV/cm, the stress response slightly increases due to the difference in electron transition probabilities.

ACKNOWLEDGMENT

The authors would like to thank Prof. Y. Kamakura of Osaka Institute of Technology for helpful discussion.

REFERENCES

[1] A. Akturk et al., *J. Appl. Phys.*, **105**, p.033703-1, 2009.
[2] T. Nishimura et al., *in Proc. of SISPAD*, p.28, 2021.
[3] BIOVIA Materials Studio DMol3 2019 User Manual.
[4] R. Fujita et al., *in Proc. of SISPAD*, p.289, 2017.

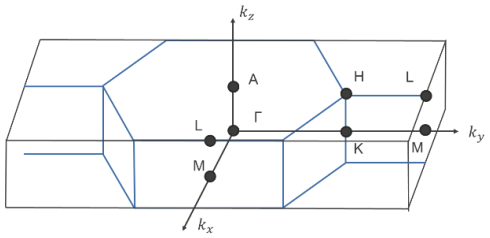


Fig. 1: Brillouin zone of 4H-SiC.

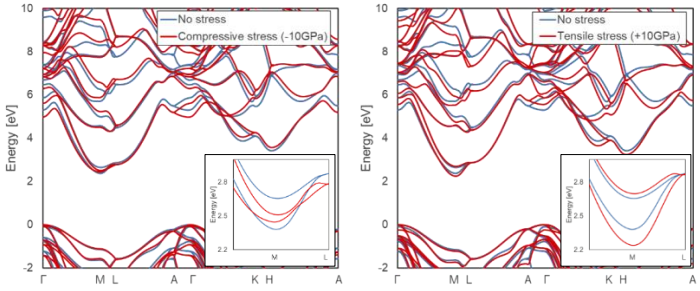


Fig. 2: Band structure change by c-axis stress.
(L) compressive stress, (R) tensile stress.

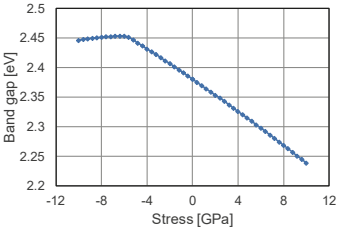


Fig. 3: Stress dependence of bandgap.

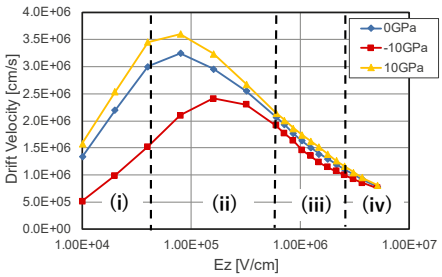


Fig. 4: Electric field dependence of drift velocity.

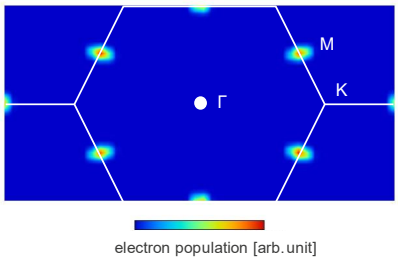


Fig. 5: Electron population in k-space @Ez=0MV/cm and 0GPa.

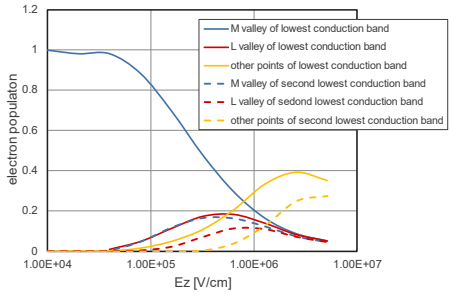


Fig. 6: Electron population @ σ =0GPa.

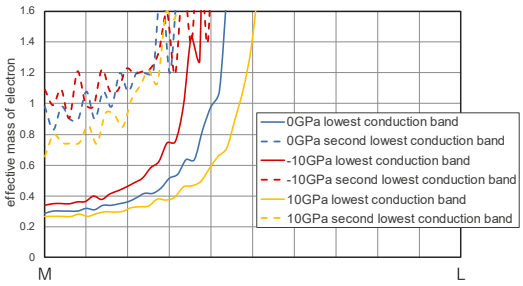


Fig. 7: Electron effective mass in M valley.

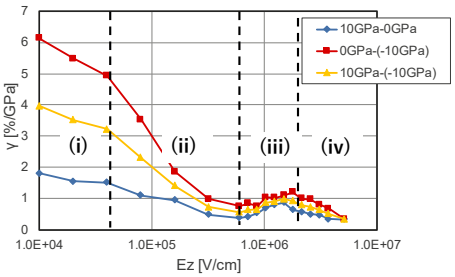


Fig. 8: Stress sensitivity of drift velocity.

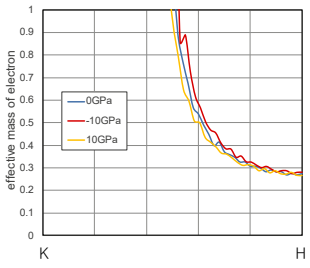


Fig. 9: Electron effective mass in H valley of the lowest conduction band.

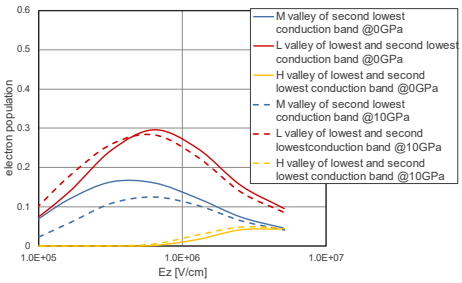


Fig. 10: Electron population around 1MV/cm.

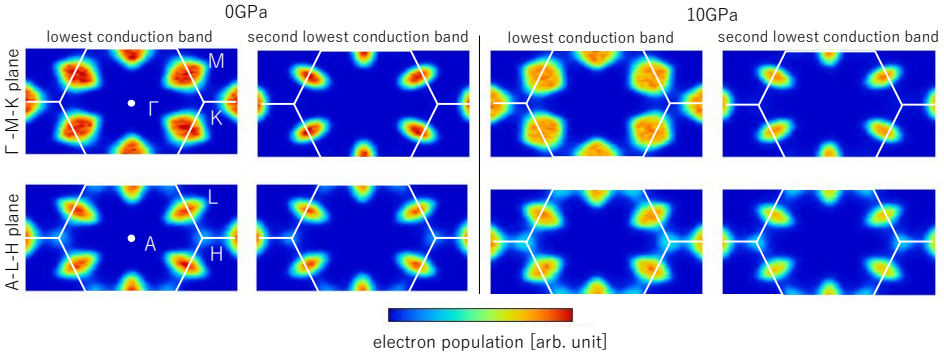


Fig. 11: Electron population in k-space @ Ez=1.28MV/cm.

Calculation of the mobility in Al_2O_3/GaN electron channel: Effect of p-doping and comparison with experiments

B. Rrustemi[1,2], F. Triozon[1], M.-A. Jaud[1], W. Vandendaele[1] and G. Ghibaudo[2]
[1]Univ. Grenoble Alpes, CEA, LETI, F-38000 Grenoble, France, Bledion.RRUSTEMI@cea.fr
[2]Univ. Grenoble Alpes, IMEP-LAHC MINATEC, F-38000 Grenoble, France

***Abstract*—In this paper, the low field electron mobility of Al_2O_3/GaN channel is calculated using a semi-classical framework. The aim is to obtain the mobility as a function of the electron sheet density and the temperature, with different p-doping conditions in GaN, and to compare it with our measurements on various samples. The scattering with bulk and surface phonons is taken into account. Scattering with dopants, interface charges, neutral impurities and interface roughness is also included in an attempt to account for experimental observations. We found that, when bringing p-dopants closer to the Al_2O_3/GaN interface, the mobility decreases mainly because the electron channel is more confined and located closer to Al_2O_3, thereby enhancing scattering mechanisms located near the oxide. Our results suggest that most of transport limitations come from Al_2O_3 oxide.**

***Index Terms*—Al_2O_3/GaN, electron mobility, p-doping.**

I. Introduction

Due to their wide band gap, high breakdown voltage and good transport properties, GaN is an attractive material for high power/frequency applications. Normally-OFF operations can be achieved with MOS-HEMT architecture which is considered promising thanks to its low leakage current [1]. In this context, investigating the transport limitations in Al_2O_3/GaN channel is important for optimizing the performances but also to feed TCAD simulations with physical models.

II. Experimental setup

GaN is grown on p-type Si substrate using MOCVD deposition. Transition layers are used for strain management and are followed by a thick and highly C-doped GaN layer (GaN:C, $[C] \approx 2.10^{19}\,cm^{-3}$) to ensure vertical insulation. p-doping is achieved with Mg. Al_2O_3 oxide is 30 nm thick and is deposited by ALD. The four samples measured are shown in Fig. 1a. Sample A has no p-doping and samples B, C and D only differ by the p-doping profile $N_A(z)$ below the Al_2O_3/GaN interface (Fig. 1b). Measurements are performed on transistors with long gate length $L_G = 60\,\mu m$ following the protocol described in [2].

III. Calculations

The electron density and potential along the stacks are obtained with self-consistent 1D Schrödinger-Poisson simulations for the Γ-valley electrons under the effective mass approximation for non-parabolic and spherical energy bands ($m_{\Gamma M} = m_{\Gamma K} \approx m_{\Gamma A}$ [3]). Non-parabolic corrections are included using Jin *et al.* method [4]. Incomplete ionization model is used to properly account for the Fermi level in GaN:C. Carbon atoms in GaN:C are assumed as deep acceptors (0.8 eV from the valence band VB) partially compensated by shallow donors (0.1 eV from the conduction band). This leads to a Fermi level pinning at 0.84 eV from the VB in GaN:C which is consistent with experimental findings [5, 6]. The p-doping is included by using $N_A(z)$ profiles (Fig. 1b) in the simulations, also accounting for incomplete ionization. The framework used to calculate the mobility is based on the linearization of the 2D Boltzmann transport equation which has been massively used in MOS inversion layers [7-9]. Scattering rates are calculated with Eq.(1) and Eq.(2) for elastic and inelastic scatterings, respectively. The mobility is then computed with Eq.(4). Screening is included in the scalar approximation [8]. The originality comes from applying this framework in Al_2O_3/GaN channel, from including neutral impurities and from emphasizing the effect of p-doping. The scattering matrix elements of acoustic deformation potential (ADP), piezoelectric potential (PE) and polar optical phonons (POP) are taken from [10]. The effect of surface optical (SO) phonons is estimated following the method of [11]. Dynamic screening has been included for POP and SO phonons as in [9] but in the scalar approximation. Fig. 2 shows the impact of screening on POP scattering and Fig. 3 compares the different phonon scatterings, highlighting the dominance of POP. Scattering with dopants, interface charges (ICs) and interface roughness (IR) is included following the approaches which have been employed in bulk MOSFET [8]. The scattering with neutral impurities (NIs) is also included: in addition to non ionized dopants, NIs are added near the Al_2O_3/GaN interface, which will yield a better agreement with experiments. To include this scattering, we use the hydrogen potential scaled by the effective mass and dielectric constant of the medium, as done by Erginsoy [12]. We account for the screening by the 2D electron gas by calculating the 2D Fourier transform of this potential (Eq.(5)), using the 3D Fourier transform given in [13]. From this potential, the screened scattering matrix elements can be computed the same way as for other mechanisms. The resulting mobility for an uniform distribution of NIs is shown in Fig. 4. At low n_s, the calculations tend to Erginsoy model but increasing n_s enhances screening and reduces the effect of background NIs, leading to an increased mobility.

IV. Comparison with experiments

The only way to account for the low effective mobility (μ_{eff}) and for the detrimental effect of p-doping on μ_{eff} (see Fig. 6) is by reinforcing scattering mechanisms located near the interface. Among those are ICs and IR scattering. A high ICs density with the same sign would induce a high V_{TH} shift. Alternative defects which do not affect the V_{TH} are NIs, hence we assume a part of the defects located near the interface to be NIs with a gaussian distribution centered at Al_2O_3/GaN interface and with a standard deviation $\sigma = 1.5$ nm. In Fig. 6, μ_{eff} vs n_s is calculated from 298 K to 423 K for the four samples and is compared to experimental measurements. With the same set of parameters, a good agreement is obtained for all samples, especially at high n_s but a high density of ICs and NIs was required to account for the low value of μ_{eff}. The ICs/NIs proportion can be modified without significantly changing the results. By bringing p-dopants closer to the interface, μ_{eff} decreases because, as shown in Fig. 5, electrons are closer to Al_2O_3, thereby increasing the effect of IR, ICs and NIs. This can be seen in Fig. 7 where the mobility limited by these scattering mechanisms is shown for 3 samples. p-doping has a huge impact on IR scattering especially as n_s decreases. The Coulomb and neutral defects scattering is mostly due to the high density of interface defects, since the scattering by p-dopants is much lower.

V. Conclusion

The mobility in Al_2O_3/GaN channel was calculated while varying the proximity of GaN p-doping to the interface. The overall good agreement between calculations and experimental data on various samples strongly suggests that the transport limitations mostly come from interface roughness and from "defects" located near Al_2O_3/GaN interface. We illustrated the effect of interface charges and neutral impurities in this study but the exact origin of these "defects", most likely due to the high-κ Al_2O_3 oxide, needs to be further investigated.

TABLE I
MATERIAL PARAMETERS USED IN THE CALCULATIONS

	E_G (eV)	m_Γ (m_0)	α (eV^{-1})	ϵ^0 (ϵ_0)	ϵ^∞ (ϵ_0)	ϵ^{int} (ϵ_0)	ω_{LO} (meV)	ω_{TO_i} (meV)
GaN[1]	3.4	0.2	0.186	10.28	5.7	-	92	ω_{TO3} =70
Al_2O_3[2]	7.0	-	-	9	3.2	7.27	-	$\omega_{TO1,TO2}$ =48, 71

[1] E_G, m^Γ from [3]. ϵ^0, ϵ^∞ from [14] and optical modes ω from [15]
[2] E_G, ϵ^0 from [16] and the rest from [11] [17].
($\epsilon^{int} \approx \epsilon^\infty (\omega_{TO2}/\omega_{LO2})^2$ as explained in [11] with ω_{LO2} taken from [17])

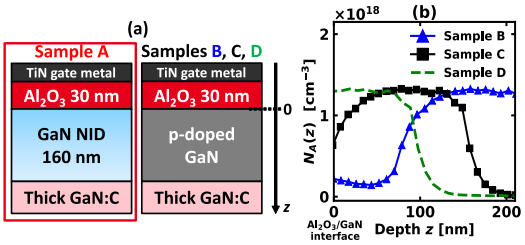


Fig. 1. (a) Stacks studied (NID: Not intentionally doped). (b) $N_A(z)$ profiles (experimental SIMS data).

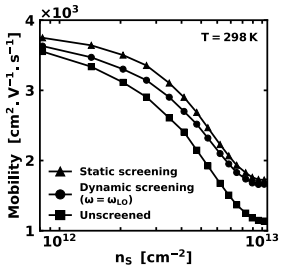


Fig. 2. POP limited mobility vs n_s. Dynamic screening is compared to the unscreened and statically screened cases.

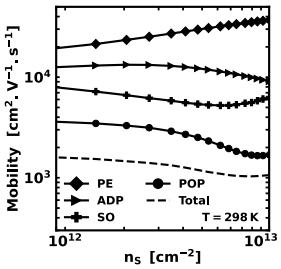


Fig. 3. All phonons limited mobility vs n_s. ADP and PE material properties are taken from [10].

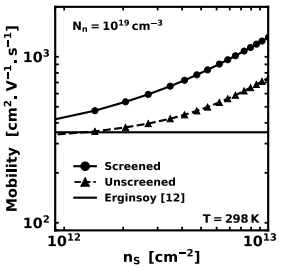


Fig. 4. Neutral impurities limited mobility vs n_s for an uniform density of neutral impurities $N_n = 10^{19}$ cm^{-3} in GaN.

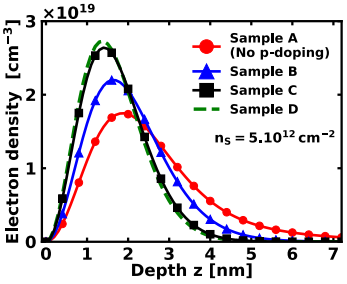


Fig. 5. Electron density vs depth z calculated for the four samples. $N_A(z)$ profiles in Fig. 1b are used in the calculations.

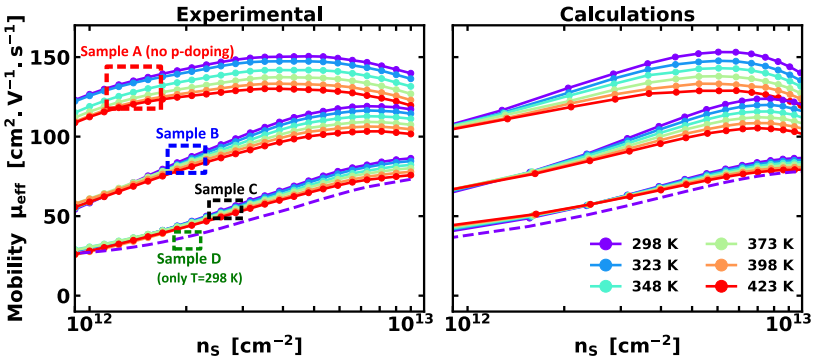


Fig. 6. μ_{eff} vs n_s from 298 K to 423 K (Experimental (left) and calculations (right)). In the calculations we used: a density of interface charges $\sigma_{Al_2O_3/GaN} = 2.1 \cdot 10^{13}$ cm^{-2}, a gaussian distribution of neutral impurities centered at Al_2O_3/GaN interface with an integrated density $N_n^s = 1.4 \cdot 10^{13}$ cm^{-2} and for interface roughness, we employed an exponential spectrum with a correlation length $\Lambda_{sr} = 1.5$ nm and a root-mean-square $\Delta_{sr} = 0.65$ nm. These parameters are not changed between samples. $N_A(z)$ profiles in Fig. 1b are used in the calculations.

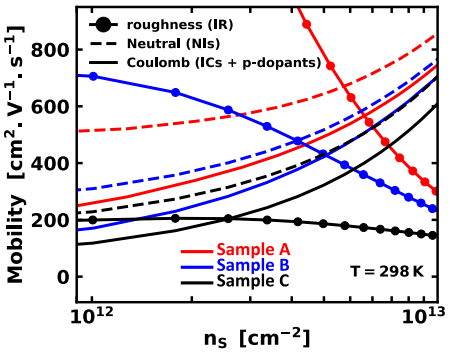


Fig. 7. Interface roughness, Coulomb and neutral impurities scattering limited mobility vs n_s for samples A, B, C. Increasing the level of p-doping and its proximity to the interface induces more confinement and enhances the effects of these scatterings. NB: the Coulomb and neutral defects scattering is mostly due to the high density of interface defects. The scattering by p-dopants (ionized and not ionized) is much lower. $N_A(z)$ profiles in Fig. 1b are used in the calculations.

$$\frac{1}{\tau_i^s(E)} = \frac{m_\Gamma}{2\pi\hbar^3} \sum_{j|E_j<E} (1+2\alpha(E-U_j)) \int_0^{2\pi} d\theta(1-cos\theta)|M_{ij}(q)|^2 \quad (1)$$

$$\frac{1}{\tau_i^{POP,SO}(E)} = \frac{m_\Gamma}{2\pi\hbar^3} \sum_{j|E_j<E\pm\hbar\omega} \frac{1-f(E\pm\hbar\omega)}{1-f(E)} (1+2\alpha(E\pm\hbar\omega-U_j)) \times \int_0^{2\pi} d\theta(1-cos\theta)|M_{ij}^{POP,SO}(q)|^2 \quad (2)$$

$$\frac{1}{\tau_i(E)} = \sum_s \frac{1}{\tau_i^s(E)} \quad (3)$$

$$\mu = \frac{-em_\Gamma}{\pi\hbar^2 n_s} \sum_i \int_{E_i}^\infty dE\tau_i(E) \frac{df^0}{dE} \frac{(E-E_i+\alpha(E-U_i)^2)}{1+2\alpha(E-U_i)} \quad (4)$$

i is the subband index, E_i the subband minimun, M_{ij} the scattering matrix elements, $q = |\mathbf{k}-\mathbf{k}'|$ the in-out wave vectors difference, α the non parabolicity factor, U_j is the expected value of the potential energy with respect to the jth subband wave function [4] and s is the scattering mechanism.

$$V_h(q,z) = \frac{-e}{2\epsilon\sqrt{4a_B^{-2}+q^2}} \left(1 + \frac{2a_B^{-2}}{4a_B^{-2}+q^2} + \frac{2a_B^{-2}}{\sqrt{4a_B^{-2}+q^2}}|z|\right) e^{-\sqrt{4a_B^{-2}+q^2}\ |z|} \quad (5)$$

V_h is the 2D Fourier transform of hydrogen potential, $a_B = (4\pi\epsilon\hbar^2)/(m_\Gamma e^2)$ is the effective Bohr radius.

REFERENCES

[1] H. Amano et al., J. Phys. D: Appl. Phys., vol. 51, no 16, p. 163001, 2018.
[2] B. Rrustemi et al., in 2021 IEEE 51st European Solid-State Device Research Conference (ESSDERC), 2021. p. 295-298.
[3] C. Bulutay et al., Phys. Rev. B., vol. 62, no 23, p. 15754, 2000.
[4] S. Jin et al., J. Appl. Phys., vol. 102, no 8, p. 083715, 2007.
[5] C. Koller et al., Appl. Phys. Lett., vol. 111, no 3, p. 032106, 2017.
[6] C. Koller et al., J. Appl. Phys., vol. 130, no 18, p. 185702, 2021.
[7] M. V. Fischetti, J. Appl. Phys., vol. 89, no 2, p. 1232-1250, 2001.
[8] D. Esseni, P. Palestri, et L. Selmi. Nanoscale MOS transistors: semi-classical transport and applications. Cambridge University Press, 2011.
[9] T. P. O'Regan et al., J. Appl. Phys., 2010, vol. 108, no 10, p. 103705.
[10] I. Berdalovic et al., J. Appl. Phys., vol. 129, no 6, p. 064303, 2021.
[11] M. V. Fischetti et al., J. Appl. Phys., vol. 90, no 9, p. 4587-4608, 2001.
[12] C. Erginsoy, Phys. Rev., vol. 79, no 6, p. 1013, 1950.
[13] T. Ouisse, J. Phys. Soc. Japan, vol. 67, no 12, p. 4157-4163, 1998.
[14] F. Bernardini et al., Phys. Rev. Lett., vol. 79, no 20, p. 3958, 1997.
[15] T. Ruf et al., Phys. Rev. Lett., vol. 86, no 5, p. 906, 2001.
[16] Z. Yatabe et al., Jpn. J. Appl. Phys., vol. 53, no 10, p. 100213, 2014.
[17] B. G. Frederick et al., Phys. Rev. B., vol. 44, no 4, p. 1880, 1991.

TCAD-based design and verification of the components of a 200 V GaN-IC platform

P. Vudumula[1], T. Cosnier[2], O. Syshchyk[2], B. Bakeroot[3], S. Decoutere[2]
[1] Department of Electrical Engineering, KU Leuven, Belgium
[2] Imec, Kapeldreef 75, 3001 Leuven, Belgium
[3] Center for Microsystems Technology, IMEC and Ghent University, 9052 Ghent, Belgium
E-mail: pavan.vudumula@imec.be

This paper describes TCAD calibration of the components processed in a 200 V GaN-on-SOI integrated circuit (IC) developed on 200 mm substrates, and verification using designed low and high voltage devices. By combining GaN-on-SOI substrates with deep trench isolation, the disclosed GaN-IC platform successfully eliminates back-gating effects in high-side power devices, as well as crosstalk between switching power devices and sensitive analog circuits [1,2]. The GaN IC platform is also supported by a variety of low-voltage analog/logic devices and passive components. Cosnier et al. [3] provides additional information about this platform in terms of device processing. The available components are presented in the schematic overviews shown in Figures 1 and 2. This platform combines monolithically integrated depletion-mode (D-mode) MIS-HEMTs and Gated-Edge-Termination Schottky barrier diodes (GET-SBDs) with an enhancement-mode (E-mode) HEMT technology baseline. The primary components of the GaN-IC platform are calibrated using a pre-existing TCAD process deck for p-GaN HEMTs as well as experimental data. Device simulations have been verified using measured low voltage test structures. Verification of simulations with the measurements results in calibration of sheet resistance (R_{sh}) in the gate and access region, threshold voltage (V_{th}), drain current (I_{ds}), ON-resistance (R_{ON}), gate current (I_g) for HEMT structures, and turn-on voltage (V_T) and forward voltage drop (V_F) for GET-SBD structure.

Partial recess of the AlGaN barrier in the gate region (process splits on remaining AlGaN barrier thickness: 6, 8, 10 nm) impacts the D-mode HEMT threshold voltage and GET-SBD turn-on voltage, while the thickness of AlGaN barrier in the access region is 12.2 nm. Figure 3 shows the TCAD calibration methodology for sheet resistance in the gate region (R_{sh_gate}) by varying the gate length of symmetrical low voltage E-mode HEMT and D-mode HEMT test structures with AlGaN recess splits, while the sheet resistance in the access region (R_{sh_access}) is 765 Ω/□ and Mobility is around 1600 $cm^2/V.s$. Figure 4 shows the capacitance-voltage (C-V)/conductance-voltage (G-V) measurement results for the calibration of interface traps (D_{it}) at the passivation/AlGaN interface [5]. Conductance values are calculated from C-V measurements of MIS capacitor for multiple frequencies in between 1 kHz to 1MHz as shown in Fig. 4a. D_{it} at the passivation/AlGaN interface calculated to be 2.5 times the peak value of conductance from Fig. 4b i.e., 6.625 x 10^{12} $eV^{-1}cm^{-2}$ with a trap energy level of 0.32 eV from conduction band (E_C-E_T). Figure 5a and Fig. 6a shows the Fermi level pinning of E-mode and D-mode HEMTs with the illustration of energy band diagrams at equilibrium condition (V_g=V_d=V_s= 0 V). Figures 5, 6, and 7 show the simulated I-V curves matching with the measured data of E-mode, D-mode HEMTs, and GET-SBDs at room temperature. Threshold voltage is extracted using maximum transconductance method and simulated V_{th} and subthreshold slope (SS) of E-mode HEMT is 2.9 V and 110 mV/dec, which matches experimental results as shown in Fig. 5b and Fig. 5c. Figure 5d shows the measured gate current in the E-mode HEMT and simulated curve by calibrating the back-to-back connection of Schottky and PIN diode in the gate region. I_d-V_g simulations of D-mode HEMT are performed to observe the impact of remaining AlGaN thickness in the gate region and dielectric thickness on V_{th}, as shown in Fig. 6b and Fig. 6c. Forward measured I-V characteristics of GET-SBDs for an AlGaN split of 9 nm are compared with simulations to observe the V_T and V_F, as shown in Fig. 7a and Fig. 7b. Table I lists the simulated parameters compared to the measured results of 200 V GET SBDs at 25° C and 150° C. In the full paper, physical models used for the simulations will be discussed in detail, along with various design parameters of high voltage devices.

[1] X. Li et al., IEEE Intern. Electr. Dev. Meet. (IEDM), 78-81, 2019, doi:10.1109/IEDM19573.2019.8993572.
[2] X. Li et al., IEEE Electr. Dev. Lett., vol. 39, no. 7, 999-1002, 2018, doi: 10.1109/LED.2018.2833883.
[3] T. Cosnier et al., IEEE Intern. Electr. Dev. Meet. (IEDM), 5.1.1-5.1.4, 2021, doi: 10.1109/IEDM19574.2021.9720591.
[4] B. Bakeroot et al., ISPSD, 419-422, 2019, doi: 10.1109/ISPSD.2019.8757629
[5] Kim, H., et al., Appl. Phys. A 126, 449, (2020), doi: 10.1007/s00339-020-03645-9

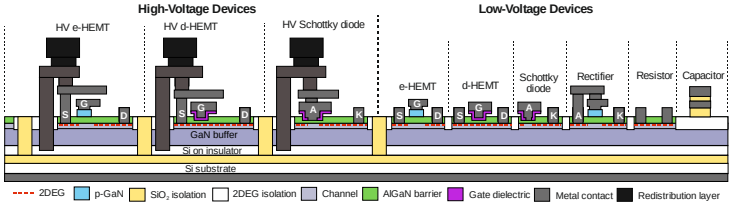


Fig. 1. Schematic cross section of IMEC 200V GaN-on-SOI Power IC technology with low and high voltage components [3].

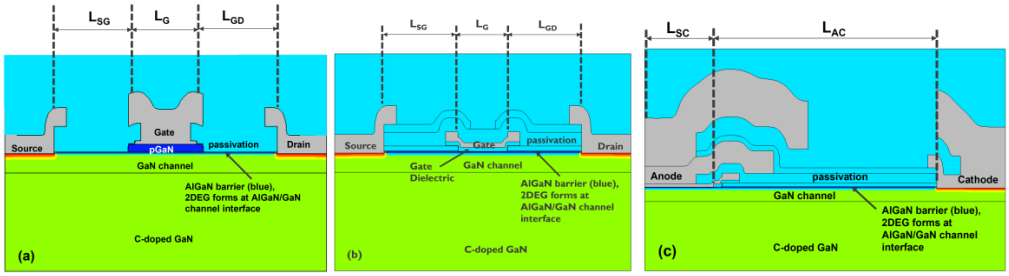
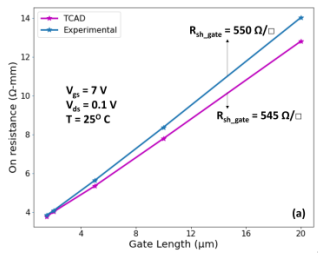


Fig. 2. TCAD calibration structures of (a) low voltage E-mode HEMT, (b) low voltage D-mode HEMT, and (c) 200 V GET-SBD with field plate configurations.



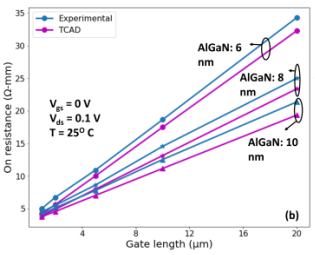
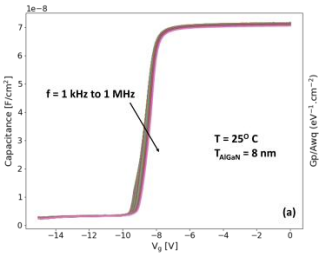


Fig. 3. Extrapolation of sheet resistance in the channel region from ON-resistance by varying the gate length of low voltage test structures (a) E-mode HEMT and (b) D-mode HEMT from 0.8 μm to 20 μm, while the sheet resistance in the access region is 765 Ω/□.



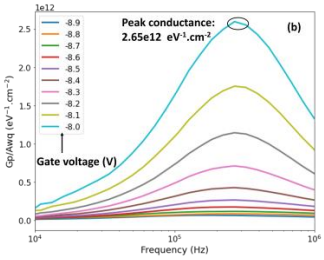


Fig. 4. Extraction of interface traps at passivation/AlGaN interface from measured (a) C-V and (b) G-V curves.

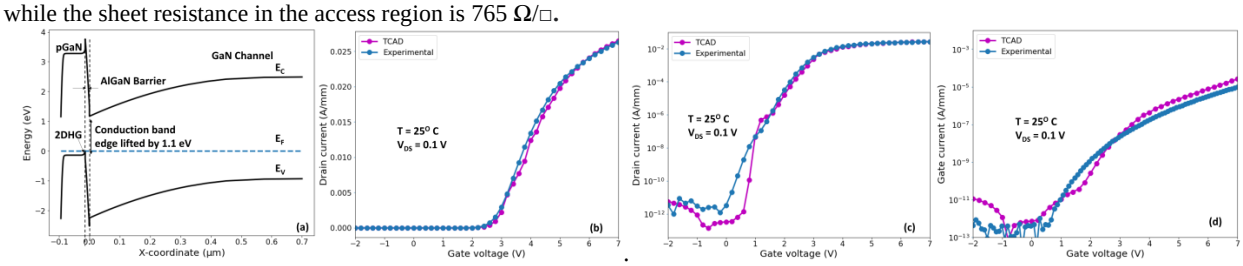


Fig. 5. (a) Band diagram of E-mode HEMT TCAD calibration structure shows unpinned Fermi level (depletion of 2DEG under pGaN at AlGaN/GaN interface) at equilibrium condition (V_g=V_d=V_s= 0 V). Transfer characteristics of low voltage test structure (L_G=L_{GD}=L_{SG}=1.5 μm) in (b) linear, (c) semi-logarithm scale, (d) gate current with the comparison of experimental results.

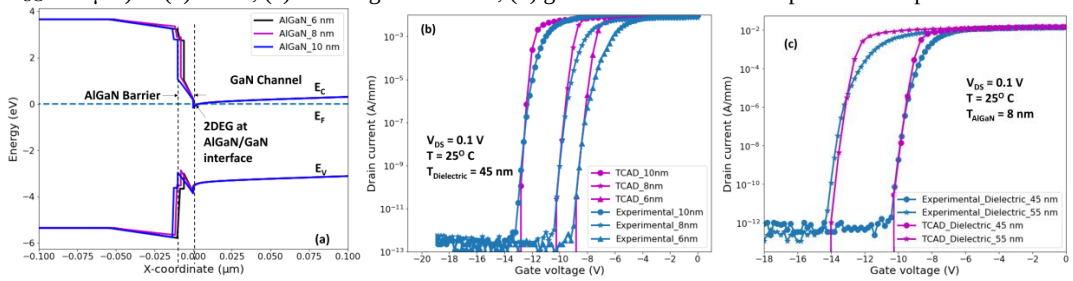


Fig. 6. (a) Band diagram of D-mode HEMT TCAD calibration structure shows the pinned Fermi level (2DEG formation) for AlGaN process splits of 6, 8, 10 nm. Transfer characteristics of low voltage test structure (L_G = 1.3 μm, L_{GD}=L_{SG}=3.0 μm) in comparison with experimental results for (b) AlGaN splits in the gate region, (c) gate dielectric thickness of 45 nm, and 55 nm.

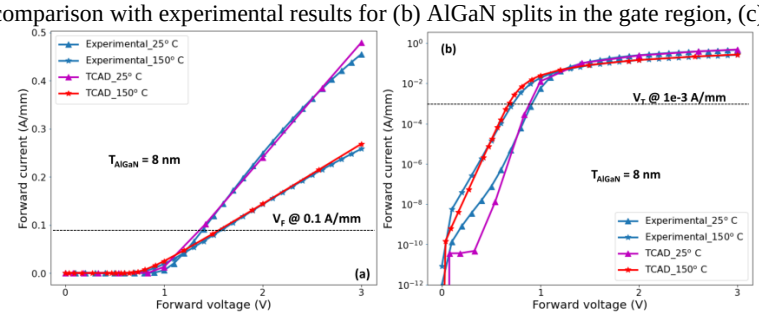


Table I: Comparison of simulated and experimental results of 200 V GET-SBD

Parameter	TCAD	Experimental
V_T_25° C	0.89 V	0.90 V
V_T_150° C	0.68 V	0.68 V
V_F_25° C	1.42 V	1.42 V
V_F_150° C	1.67 V	1.67 V
Ideality factor η_25° C	1.82	2.06
η_150° C	1.21	1.47

Fig. 7. Forward I-V characteristics of 200 V GET SBD (L_{SC} = 2 μm, L_{AC} = 6.5 μm) in (a) linear and (b) semi-logarithm scale with the comparison of experimental results.

Development of an Ensemble Monte Carlo Simulator for High-Power Semiconductor Devices with Self-Consistent Electromagnetism and GPU Implementation

S.J. Cooke[+] and M.G. Ancona*[+]

[+]Electronics Science and Technology Division, Naval Research Laboratory
Washington, DC 20375, USA simon.cooke@nrl.navy.mil

*Department of Electrical and Computer Engineering, Florida State University
Tallahassee, FL 32310, USA

The ensemble Monte Carlo (EMC) method is a powerful and widely used technique for modeling carrier transport in semiconductors in which one can easily take into account band structure and a variety of scattering mechanisms. Because it incorporates the microscopic physics, it can be particularly useful for predicting the potential of new semiconductor materials. In the usual implementations of EMC, the semi-classical electron dynamics is coupled self-consistently to the electrostatic field and this suffices for most devices. However, for high power, high current semiconductor device applications the electrostatic approximation is often not adequate and the full electromagnetic equations (EM) must be solved in fully-coupled fashion. In this paper we report on the development of a fully self-consistent, time-domain, EM-EMC simulator in which we also address a well known drawback of these methods, their computational intensiveness, by producing a GPU implementation.

High-power, high-current semiconductor devices, such as fast opening switches or high-power amplifiers, often exhibit effects such as strong magnetic fields, filamentation and instability that are more often associated with vacuum plasmas [1]. With this in mind, a convenient starting point for our development of a particle-based simulator of these solid-state plasma processes is an existing 3D electromagnetic particle-in-cell (EM-PIC) simulation framework called Neptune [2], originally created for modeling high-power, relativistic, vacuum electronic devices. This code already includes a GPU implementation for computational efficiency.

Solving Maxwell's equations in place of Poisson's equation introduces a number of challenges. While Poisson's equation is defined in terms of electric potential and charge density distributions, Maxwell's time-dependent equations fundamentally use electromagnetic fields and currents. Neptune's EM-PIC algorithm preserves the continuity equation exactly so that Poisson's equation remains satisfied implicitly for the duration of the simulation. The character of the field equation solution has also changed from elliptic to hyperbolic. The time integration of electromagnetic fields in Neptune uses an explicit leapfrog scheme related to the Finite Difference Time Domain (FDTD) method, for which the maximum time step is constrained by the Courant condition for stability. For typical dimensions and particle velocities found in semiconductor simulations this leads to time steps much shorter than when using a Poisson solver, but at a lower computational cost per time step. In the present implementation, the particle time step is taken to be the same as the electromagnetic time step and, because of this, the implementation of the scattering process in our model differs from that in conventional EMC implementations. Rather than computing a time-of-flight until the next scattering event takes place, the probabilities of each scattering mechanism are evaluated independently each time step.

The primary adaptations to the EM-PIC code to implement the fully-coupled EM-EMC model lie in the calculation of the time-step updates to the momentum and position for each

particle, which in Neptune are computed in parallel on the GPU. Firstly, we modified the energy-momentum relation (which is used to compute the particle velocity) from the hyperbolic relativistic electron formula to a non-parabolic semiconductor band-structure model. Secondly, we introduced a set of probabilistic scattering mechanisms to represent phonon scattering events. Thirdly, we implemented the calculation of random thermal distributions of particles' energies and momenta that are required for the initialization, injection, and scattering of particles.

Each simulation begins in a charge-neutral state with no applied voltages, initialized with electromagnetic fields set to zero and an ensemble of simulation particles distributed according to the donor density distribution. Voltages are introduced by applying controlled current sources between electrodes to transfer charge and generate electric fields in the device, which interact self-consistently with the simulation particle ensemble until a steady-state solution is reached. Simulation particles are added and removed at the electrodes during the simulation to model Ohmic contacts.

To illustrate the EM-EMC simulator in 2D initially, we applied it to a quasi-1D n^+-n-n^+ diode and to a Si MESFET. Figure 1 illustrates the longitudinal electric field and particle ensemble distributions across a n^+-n-n^+ Si diode at steady state with 1V applied between the end electrodes. Figure 2 illustrates the computed electric field magnitude and particle distributions at steady state inside a Si MESFET with 1V applied between the source and drain electrodes. Full details of the method, simulation results, and simulation times will be presented at the meeting.

Acknowledgement: The authors thank the Office of Naval Research for funding support.

1. P. M. Platzman and P.A. Wolff, *Waves and Interactions in Solid State Plasmas*, (Academic Press, First Edition 1973)
2. S. J. Cooke, I. A. Chernyavskiy, G. M. Stanchev, B. Levush, T. M. Antonsen Jr. "GPU-Accelerated Large-Signal Device Simulation Using the 3D Particle-in-Cell Code 'Neptune'" 39th IEEE International Conference on Plasma Science, Edinburgh, UK, July 8-12, 2012.

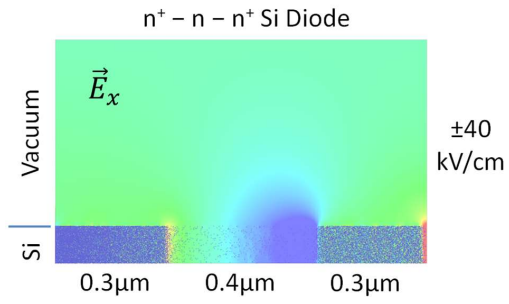


Figure 1. Longitudinal electric field strength in a quasi-1D n+−n−n+ Si diode after steady-state operation is reached (red is +ve, green is zero, blue is –ve).

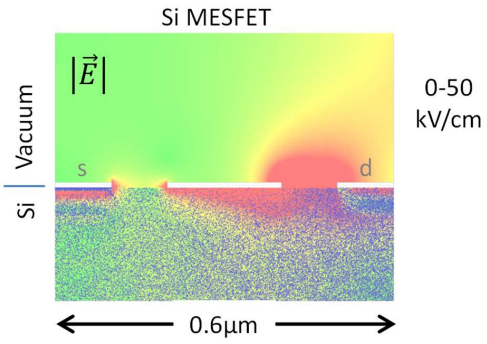


Figure 2. Electric field magnitude in a 2D Si MESFET simulation after steady-state operation is reached.

Investigation of effects of lateral boundary conditions on current filament movements in Trench-Gate IGBTs

Takeshi Suwa
Toshiba Electronic Devices & Storage Corporation, Kawasaki, Japan, Email: takeshi.suwa@glb.toshiba.co.jp

When designing the reliability of high voltage power devices, it is very useful to confirm the behavior of current filaments in various structures and situations by TCAD simulation [1]. Current filaments are caused by the instability inherent in the device, which is the concentration of currents in only some cells and moves around in the device. In general, current filaments move around in the chip plane and three-dimensional (3D) simulations are appropriate. For simplicity, device simulations are often performed with a structure that limits one direction of current filament movements, that is, a 2D cross section in which the chip is cut vertically. Even in the 3D case, it is common to cut out a part of the chip as a simulation structure and set boundary conditions on each cut out surface. Which structure we use for the simulation depends on our purpose and available time. In any case, for current filament simulations, the lateral boundary conditions are especially important as the filaments move laterally. It is known that when reflection boundary conditions (ideal Neumann) are used in 2D current filament simulations, the boundary conditions can significantly affect the results [2]. This time, I report the effect of the lateral boundary conditions of a 3D structure on the current filament movements and heat generation.

Approach

In this study, overcurrent turn-off phenomena are simulated by Synopsys TCAD with the thermodynamic model. As shown in Fig. 1, the simulated structure consists of 8 cells in the X direction and 16 cells in the Z direction, with 4 cut out surfaces on the lateral sides. Reflective or (Mortar) periodic boundary conditions are imposed on these lateral surfaces in three ways (a), (b) and (c). Especially in 3D calculations with the iterative solver, if the periodic boundary condition is imposed, the convergence becomes very poor, so it is necessary to carefully set the mesh generation and use a specialized linear solver for the mortar periodic boundary condition (Schur Solver implemented in SDevice).

Results

The results of the overcurrent turn-off simulation under the three boundary conditions are shown in Fig. 2. As can be seen from the figure, the waveforms of the collector-emitter voltage Vce, the gate-emitter voltage Vge and Tmax start to be jagged when the collector current Ic starts to decrease. The jaggedness of these waveforms is caused by the generation and the movement of a high current density current filament. The main cause of the current filament movements is that the lattice temperature becomes locally non-uniform due to Joule heat, and the filament moves to lower temperature parts where impact ionization is likely to occur. As can be seen from the jaggedness in the figure, the heat generation inside the device differs greatly depending on how the boundary conditions are imposed.

Fig. 3 shows the current density and the lattice temperature distributions under the boundary condition (a) at each current value. The 2D distributions cut out horizontally on the cut plane are also shown below. In the 2D current density distributions, some arrows show the trajectory of the current filament movements. The current filament initially crosses the center of the device and hits the Xmin plane. After that, it moves along the boundaries and disappears as Ic decreases. In the case of the simulation with a 2D structure, there is no place to escape, so the filament bounces off the side boundary line, but in a 3D case it travels along the boundary surface. The reason why the filament moves along the boundaries and the filament occurs at the boundary surface is that under the reflection boundary condition, the values of the physical quantities near the boundary surface become large when there are physical quantity flows in the horizontal direction. For this reason, the current filament becomes difficult to move due to the hole injection from the back surface and the leakage current from the front surface. The increase in the leakage current is due to temperature rise and potential barrier drop in the base region. After all, the current filament is difficult to move due to the influence of the reflection boundary condition. The difficulty of movements is also reflected in the large jaggedness of the waveform shown in Fig. 2.

Fig. 4 shows the same distributions as in Fig. 3 under the boundary conditions of (b) and (c). In (c), the timing of a current filament formation is the latest and the filament does not move along the boundary. Due to the periodic boundary condition, the filament that reach Xmin plane emerges from Xmax plane. After that, it is scattered by the high heat part formed at the position where it was first generated, and it moves toward Zmax and disappears as Ic decreases. In (b), it can be seen that the movement is as if the features of both (a) and (c) are combined. When looking at the effect of the current filament moving through the cell, it is more appropriate to model the phenomenon under periodic boundary conditions.

References [1] T. Tamaki et al., "Numerical study of destruction phenomena for punch-through IGBTs under unclamped inductive switching," Microelectronics Reliability, Vol. 64, September 2016, pp. 469-473.

[2] G. De Falco et al., "Effects of current filaments during dynamic avalanche on the collector-emitter-voltage of high voltage Trench-IGBTs," EPE'15 ECCE-Europe.

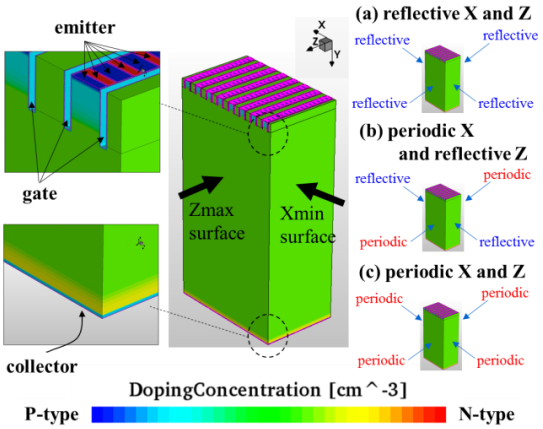


Fig. 1: Schematic view of the simulated structure (Trench-Gate IGBT 8x16 cells). (a), (b), and (c) represent three types of boundary condition settings for the cut out vertical plane.

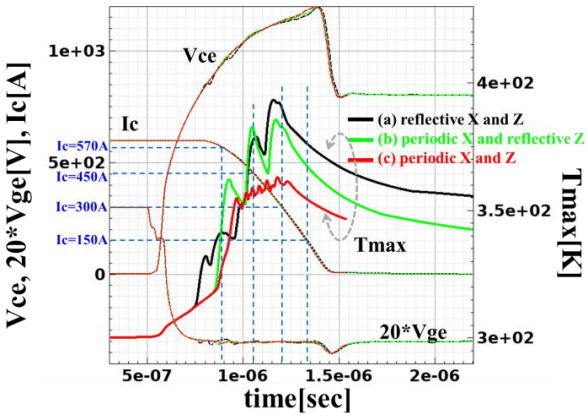


Fig. 2: Waveforms of overcurrent turn-off simulation calculated under the three boundary conditions in Fig. 1. (c) is now being calculated. Tmax represents the maximum lattice temperature of the Si region at each time.

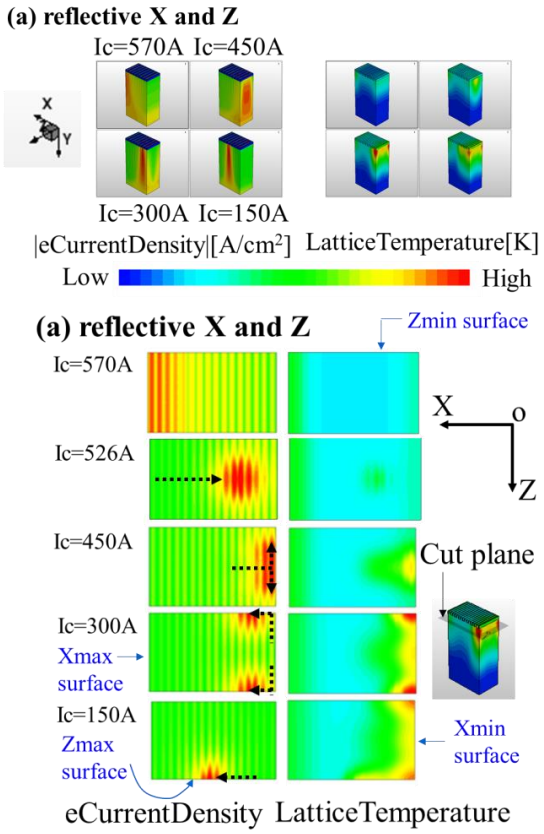


Fig. 3: Distributions of electron current density and lattice temperature at each collector current value in the case of (a) in Fig. 1. The lower figures show the distributions on the horizontal plane cut out by "Cut plane" in the figure. The arrows in the figure indicate the trajectory of filament movement.

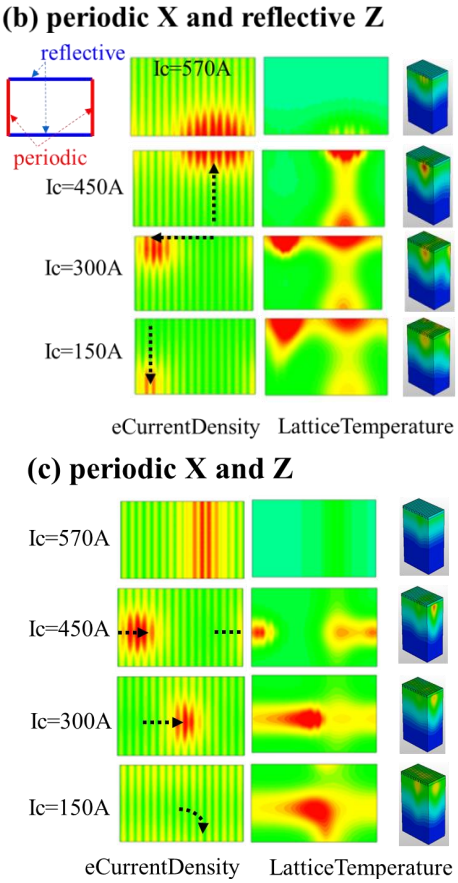


Fig. 4: Distributions similar to Fig. 3 in the case of (b) and (c). At 150A in (c), the current filament disappears, but the temperature distribution shows the movement.

Assessment of Lateral and Vertical Tunneling FETs Based on 2D Material for Ultra-Low Power Logic Applications

Yuanchen Chu[1,3], Shang-Chun Lu[2,3], Michael Povolotskyi[1], Gerhard Klimeck[1], Umberto Ravaioli[2], Tomás Palacios[3], and Mohamed Mohamed[4]

[1]*School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907,* [2]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801,* [3]*Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA 02139,* [4]*MIT Lincoln Laboratory, Lexington, MA 02421; email: mohamed.mohamed@ll.mit.edu*

Introduction: The relentless push for scaling has led to successfully packing over a billion transistors into a single chip, and such dense packing has resulted in increased functionality and performance of electronic devices that permeate our everyday life. However, the march to transistor miniaturization is stumbling over technical and physical barriers. As we reach the scaling limit of CMOS logic, it is imperative to assess alternative energy-efficient electrical switching elements. Steep subthreshold devices—with room temperature subthreshold swing (SS) below the 60mV/dec thermal limit– are desirable as they allow voltage scaling and hence enable power reduction and energy-efficient computing. Tunneling Field Effect Transistor (TFET) is arguably the most mature and promising low-power steep SS post-CMOS device, with numerous reports experimentally demonstrating sub-thermal SS [1]. However, previous TFET efforts with sub-thermal SS and acceptable I_{OFF} were plagued with poor on-current (I_{ON}) and high interfacial trap density (D_{it}). The emergence of two-dimensional (2D) materials promises to resolve these issues, as 2D materials facilitate bandgap tuning and an ideal trap-free interface. This work uses ab-initio quantum simulation to examine the promise and limitations of homojunction transition metal dichalcogenides (TMDCs) and black phosphorus (BP) and explores the potential of 2D/2D vertical heterojunction TFET design for boosting drive current.

Simulation Approach: We adopt an ab-initio quantum simulation framework utilizing first-principle density functional theory (DFT) to extract material properties and non-equilibrium Green's function (NEGF) for electron transport. The VASP DFT tool is employed to calculate the electronic structures of 2D materials [2]. The Bloch functions provided by DFT are then transformed into maximally localized Wannier functions (MLWFs) using the wannier90 package [3]. This process produces a first-principle tight-binding (TB) like Hamiltonian of the unit cell, which, together with the unit cell atom positions and lattice vectors, are loaded into the NEMO5 NEGF quantum solver for electrical characterization [4]. For all devices, flat band Neumann boundary conditions are assumed for the Poisson equation, except for the gate/oxide interface regions where Dirichlet conditions are used to set the applied gate voltage.

Preliminary Device Results: Fig. 1 (a)-(b) show the top-and side-view of BP and TMDC crystals. Fig. 1 (c) provides the basic device schematic for homojunction 2D TFET. Comparisons of the DFT and MLWF band structure of 1L WTe_2 and 1L/2L/3L BP are shown in Fig. 2. The results show good transferability of the parameter set used in this work. The bandgap of 1L/2L/3L BP and TDMC agree well with previous calculations and experiments. Figure 3 summarizes transport results using the hybrid DFT/NEGF approach in which the Kohn-Sham orbitals over Bloch-wave vectors provided by DFT are transformed into TB-like orbitals based on MLWF. The IV characteristics of monolayer 2D TFETs are shown in Fig. 3a. The TFET I_{ON} trends for the 2D materials is BP > WTe_2 > WSe_2 > $MoTe_2$ > WS_2 > $MoSe_2$ > MoS_2. BP and WTe_2 are the only two materials with $I_{ON} \geq 100$ μA/μm due to their favorable material properties. In Fig. 3(b), we examine the difference between the two best monolayer TFET devices by plotting the complex band structure. The complex band structure is related to the evanescent wave function, $\exp(-\kappa x)$, where $i\kappa$ is the imaginary wave vector. The area enclosed by the imaginary wave-vector and the energy axis is related to the band-to-band tunneling (BTBT) probability decay rate. BP's higher I_{ON} is attributed to the slower decay of the BTBT probability in the transverse reciprocal k-space. Lastly, we explore new device designs based on 2D/2D vertical heterojunction to achieve a higher drive current suitable for VLSI applications. We propose a vertical device design with a broken bandgap, as shown in Fig. 4, which achieves increased I_{ON} values than monolayer devices. We will discuss the vertical device operation principles and examine the impact of band tails on device performance.

References: [1] D. Sarkar et al. Nature vol. 526, p91–95 (2015), [2] G. Kresse *et al., Phys. Rev. B,* vol. 54, p. 11169, (1996), [3] A. A. Mostofi *et al., Comput. Phys. Commun,* vol. 178, p. 685, (2008), [4] S.Steiger *et al., IEEE Trans. Nanotechnol.*,vol. 10, p. 1464 (2011)

Acknowledgment: This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

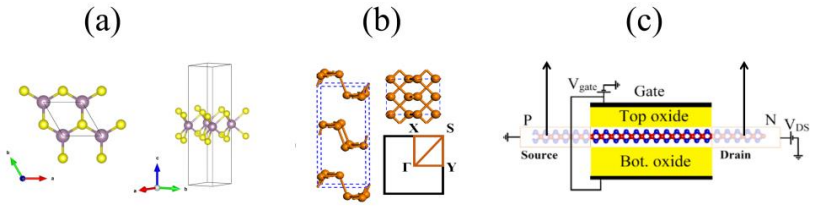


Fig. 1 (a) top view and side view (left) of the TMDC crystal structure. Yellow balls represent chalcogens and light purple balls represent transition metals (TM). (b) Top view and side view of BP crystal structure. (c) Device structure of monolayer TDMC

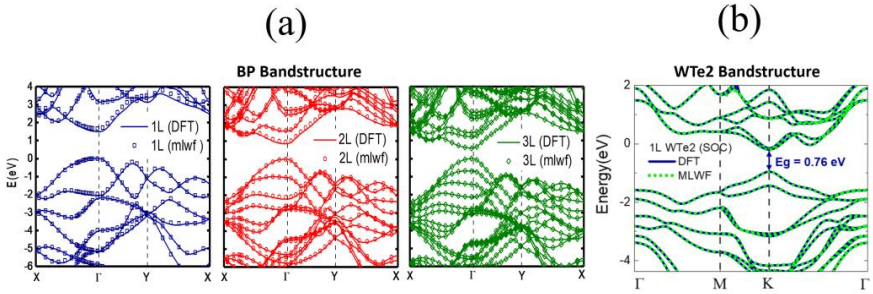


Fig. 2 DFT vs. MLWF band structure for (a) Monolayer (1L), bilayer (2L), and trilayer (3L) BP. (b) Monolayer TDMC band structure example for WTe_2 showing transferability and reproducibility of DFT band using MLWF.

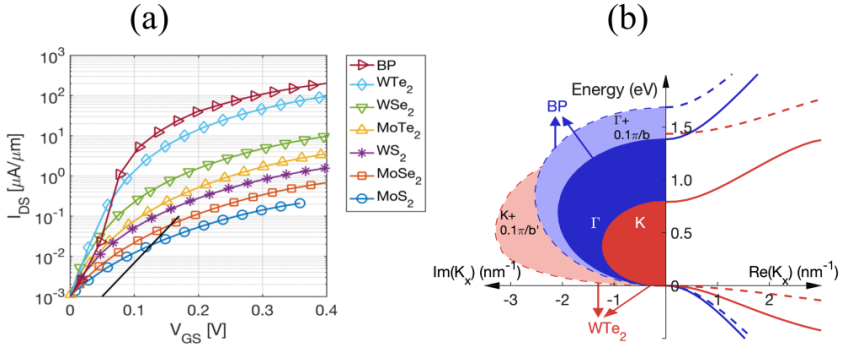


Fig. 3 (a) IV characteristics of monolayer TFETs. (b) Energy versus real part $Re(K_x)$ and imaginary part $Im(K_x)$ of the wave vector along the channel for BP and WTe_2.

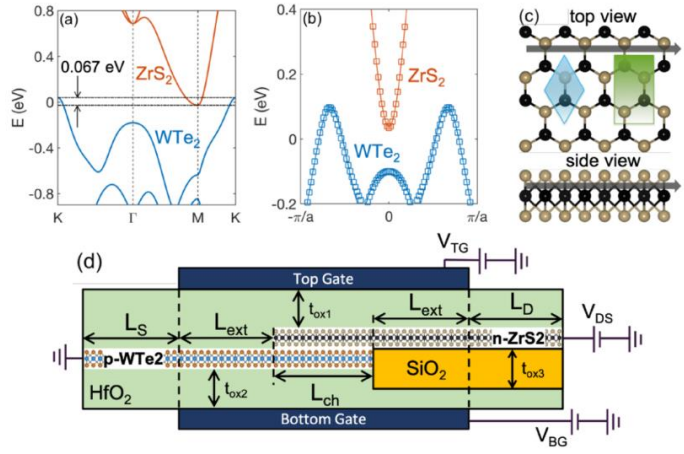


Fig. 4 (a-b) Vertical TFET design showing band structure calculated with VASP (lines) and NEMO5 (symbols). The NEMO5 band structure shown in (b) is computed along the transport direction. (c) The primitive unit cell (blue) and the rectangular unit cell (green). The vertical TFET device design is shown in (d).

Full Quantum Simulation of Shockley-Read-Hall Recombination in p-i-n and Tunnel Diodes

A. Pilotto, P. Dollfus, J. Saint-Martin, M. Pala
Université Paris-Saclay, CNRS, C2N, 10 Boulevard Thomas Gobert, 91120 Palaiseau, France

Introduction. Thermal generation-recombination mechanisms occurring in the active region of electron devices are usually a limiting factor for the performance. For instance, in single photon avalanche diodes (SPADs) they affect both the *dark count rate* and the *after pulsing*, while in Esaki diodes they are responsible for the increase of the *valley current*. Recently, a model to include Shockey-Read-Hall (SRH) type recombination [1] via multiphonon relaxation [2] in NEGF calculations has been proposed in [3], and successfully applied to the study of forward biased p-i-n junctions [3] and LEDs [4]. In this paper, we use the model of [3] to study the impact of SRH recombination occurring at different device regions on the *forward and reverse* current-voltage characteristics ($I-V$) of GaAs p-i-n junctions and of an InGaAs Esaki diode.

Model. The model for the treatment of SRH recombination via multiphonon relaxation in NEGF calculations [3] assumes that the defect density is strongly localized in space and energy. In the case of a p-i-n diode, the density of states is therefore represented by a Dirac's delta at $z{=}z_{def}$ and $E{=}E_{def}$. Here, as in [3, 4], the Dyson's equation for the retarded Green's function of the defect (G_d) is not solved, but, instead, $G_d^<$ and $G_d^>$ are computed by using the quasi-equilibrium expressions $G_d^<(E){=}if_d(E)A_d(E)$, and $G_d^>(E){=}-i(1-f_d(E))A_d(E)$, where A_d is the defect's spectral function and f_d describes the defect's occupation. The value of $f_d(E{=}E_{def})$ is given by the balance at steady-state between the capture and emission rates [3]. In the case of electrons, the self-energies that describe carrier capture/emission into/from the defect state and into/from the bands are computed with Eqs. (1) and (2), respectively.

$$\Sigma_{dc}^{\lessgtr}(z_{def},E){=}\frac{1}{\mathcal{A}}\sum_{\mathbf{k}_\parallel}\sum_{l\geq 0}\mathcal{M}_{em/capt}(l) \times\int dz G_c^{\lessgtr}(\mathbf{k}_\parallel,z,z,E+l\hbar\Omega_0)\delta(z-z_{def}) \quad (1)$$

$$\Sigma_{cd}^{\lessgtr}(z,z,E){=}\rho_d\sum_{l\geq 0}\mathcal{M}_{em/capt}(l)G_d^{\lessgtr}(E-l\hbar\Omega_0)\delta(z-z_{def}) \quad (2)$$

$$\mathcal{M}_{capt}(l){=}\mathcal{M}_{dc}^0\frac{(l-S)^2}{S}e^{-S(2f_{BE}+1)}\left(\frac{f_{BE}+1}{f_{BE}}\right)^{l/2}\mathcal{I}_l(c) \quad (3)$$

where ρ_d is the defect density, Ω_0 is the phonon energy (36 meV and 34 meV for GaAs and InGaAs, respectively), l identifies the number of phonons involved in the recombination process, S is the Huang-Rhys factor (3.5 for GaAs [2] and 0.4 for InGaAs [5], respectively), f_{BE} is the phonon occupation, $\mathcal{I}_l$ is the modified Bessel function of l-th order, $c{=}2S\sqrt{f_{BE}(f_{BE}+1)}$, and $\mathcal{M}_{em}(l){=}\mathcal{M}_{capt}e^{-l\hbar\Omega_0/k_BT}$. The constant $\mathcal{M}_{dc}^0$ can be linked to the carrier thermal velocities ($v_{th,n/p}$) and capture cross sections ($\sigma_{n/p}$) as described in [3]. Similar expressions hold for holes. Current conservation is obtained by self-consistently solving the equations for the retarded and the lesser(greater)-than Green's functions.

Bulk GaAs. As a validation of our implementation, we have verified that the recombination rate computed by using the aforementioned approach (Eq. (4)) is equivalent to the semiclassical expression (Eq. (5)) [1] when the the carrier lifetimes are $\tau_{n/p}{=}\rho_d\sigma_{n/p}v_{th,n/p}$ and the electron and hole densities are obtained from NEGF calculations. The results in Fig. 1, obtained by using an effective mass Hamiltonian with two parabolic bands, and for a defect density $\rho_d{=}10^{14}$ cm^{-3} show an excellent agreement between Eqs. (4) and (5).

$$U_{SRH}{=}\rho_d\int\frac{dE}{2\pi\hbar}\left[\Sigma_{dc}^<(E)G_d^>(E)-\Sigma_{dc}^>(E)G_d^<(E)\right] \quad (4) \qquad U_{SRH}{=}\frac{np-n_{eq}p_{eq}}{\tau_p(n+n_1)+\tau_n(p+p_1)} \quad (5)$$

GaAs p-i-n Diodes. First, we have employed the model to simulate the recombination current characteristics of GaAs p-i-n diodes with different thickness of the intrinsic region ($L=40$ nm and $L=80$ nm). A doping concentration $N_d=N_a=10^{17}$ cm^{-3} has been assumed in the contact regions and a linear potential profile across the intrinsic layer has been considered. For the two diodes, five different traps configurations have been analyzed: a single midgap defect located at $z_{def}=L/4$, $L/2$, or $3L/4$ and two random distributions of five traps, namely $R1_{40}=[4\text{ nm}, 7.6\text{ nm}, 10\text{ nm}, 23.3\text{ nm}, 39.2\text{ nm}]$, $R2_{40}=[10\text{ nm}, 15\text{ nm}, 20\text{ nm}, 25\text{ nm}, 30\text{ nm}]$ for the 40 nm thick diode and $R1_{80}=[10\text{ nm}, 11.6\text{ nm}, 15.2\text{ nm}, 32.8\text{ nm}, 57.2\text{ nm}]$, $R2_{80}=[20\text{ nm}, 30\text{ nm}, 40\text{ nm}, 50\text{ nm}, 60\text{ nm}]$ for the longer diode. Results are shown in Fig. 2. As expected, in forward bias the recombination current increases with ideality factors that approach or exceed 2 (see Fig. 3). Ideality factors larger than 2 in the case of large built-in fields (i. e. short diodes) arise from the involvement of band tail states in the trap assisted tunneling process [3]. On the other hand, in reverse bias configurations the recombination current is slowly varying with the applied bias, except for those cases where a single defect is located far from the center of the intrinsic region. In fact, by looking at the spectral current densities in Fig. 4, we notice that, due to the position of the defect with respect to the Fermi levels of the contacts, trap assisted tunneling through traps located at the center of the intrinsic region is the one that mostly affects the recombination current, while recombination at other locations becomes significant only for large applied bias voltages.

InGaAs Esaki Diode. Finally, we have simulated a typical InGaAs Esaki diode composed of two degenerate contact regions ($N_a{=}3{\times}10^{19}$ cm^{-3}, $N_d{=}10^{19}$ cm^{-3}) separated by a 3 nm thick undoped layer. A two-bands $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian has been used to simulate the coupling between the VB and the CB [6]. The coupling term has been treated as an adjustable parameter and it has been chosen in order to yield the correct curvature of the bands at the minimum/maximum. The potential profile has been derived from the self-consistent solution of the Poisson-NEGF equations. Figure 6 compares the diode's $I-V$ characteristics with and without the SRH recombination. We notice that the inclusion of recombination in the simulations translates into a higher current peak, a larger lobe of the $I-V$ curve and also into a larger valley current. The effect of SRH recombination is illustrated by Fig. 6: while at $V{=}0.1$ V (i.e. close to the peak-current voltage) both the recombination and the tunneling component contribute to the current density, only the former one determines the valley current.

Conclusions. We have employed a quantum model to include SRH recombination in NEGF simulations. The model has been used to study how the defect location in the active region of p-i-n diodes affects the characteristics of these devices, showing that defects located at the center of the intrinsic regions are the ones that are mostly interested by trap assisted tunneling. In the simulation of Esaki diodes, the inclusion of the SRH recombination is able to reproduce realistic valley currents with a reduction of the peak-to-valley current ratio from 3400 to 22. Therefore, this model can shed light on the role played by traps in the current characteristics of low-power devices such as the tunnel-FET.

[1] W. Shockley et al., Phys. Rev. 87, 835 (1952).
[2] A. Schenk, J. Appl. Phys. 71, 3339 (1992).
[3] U. Aeberhard, Phys. Rev. B 99, 125302 (2019).
[4] J. A. G. Montoya et al., Phys. Rev. Appl. 16, 044023 (2021).
[5] A. Schenk et al., ECS Transactions 66(5), p. 157 (2015).
[6] E. O. Kane, J. of Phys. and Chem. of Solids 12(2), p. 181 (1960).

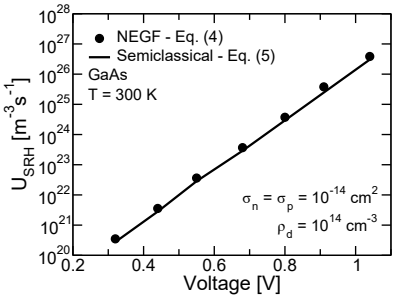


Figure 1: Recombination rate in bulk GaAs as a function of the applied bias voltage. Results obtained by using the NEGF formalism (Eq. (4), symbols) are compared to the semiclassical expression [1] (Eq. (5), lines) for σ_n=σ_p=10^{-14} cm^2 and ρ_d=10^{14} cm^{-3}.

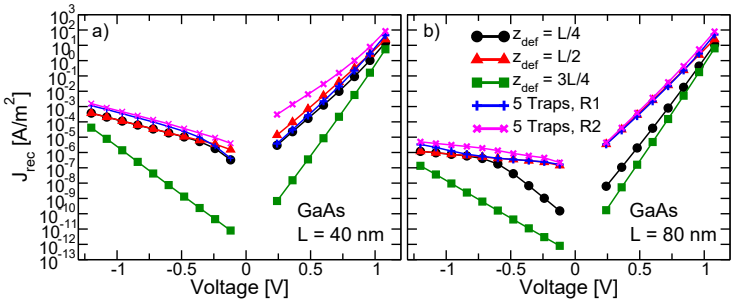


Figure 2: Recombination current as a function of the applied voltage for GaAs p-i-n diodes with intrinsic regions thickness (a) L=40 nm and (b) 80 nm. Different configurations have been considered: single midgap defects located at $L/4$ (black circles), $L/2$ (red triangles), $3L/4$ (green squares) and two random distributions of five traps (blue plus signs and magenta crosses). σ_n=σ_p=5×10^{-14} cm^2 and ρ_d=10^{13} cm^{-3}.

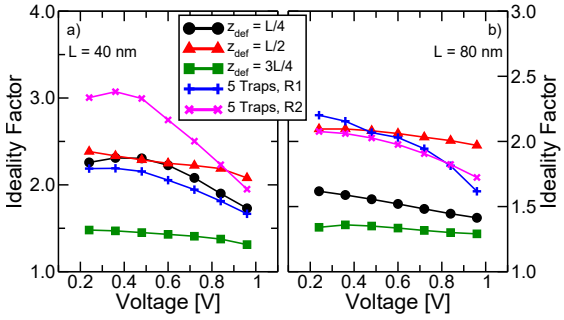


Figure 3: Ideality factors in forward bias for the p-i-n diodes and the trap configurations of Fig. 2.

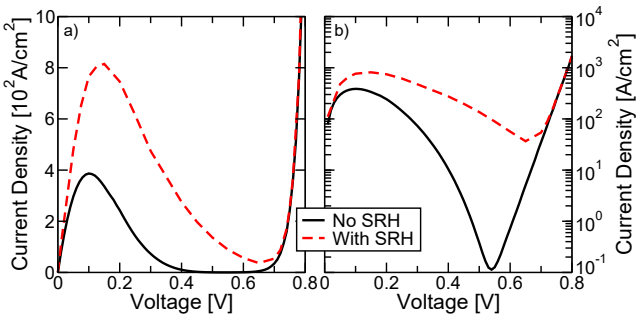


Figure 5: $I-V$ characteristics in (a) linear and (b) log scales of an InGaAs Esaki diode when SRH recombination is neglected (solid black line) or included (dashed red line). σ_n=σ_p=10^{-15} cm^2 and ρ_d=2×10^{11} cm^{-3}.

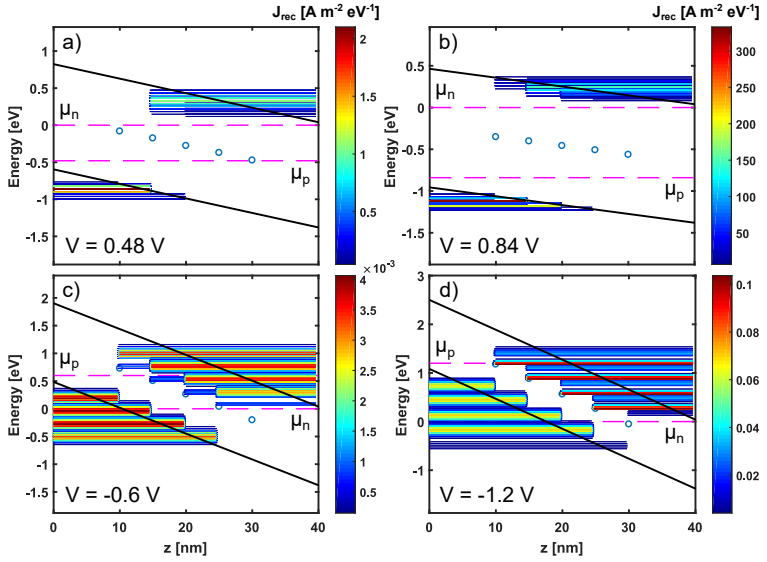


Figure 4: Spectral recombination current density at (a) V=0.48 V, (b) V=0.84 V, (c) V= − 0.6 V, and (d) V= − 1.2 V of a 40 nm long GaAs p-i-n diode when five midgap defects from z=10 nm to z=30 nm are simulated (blue circles). σ_n=σ_p=5×10^{-14} cm^2 and ρ_d=10^{13} cm^{-3}.

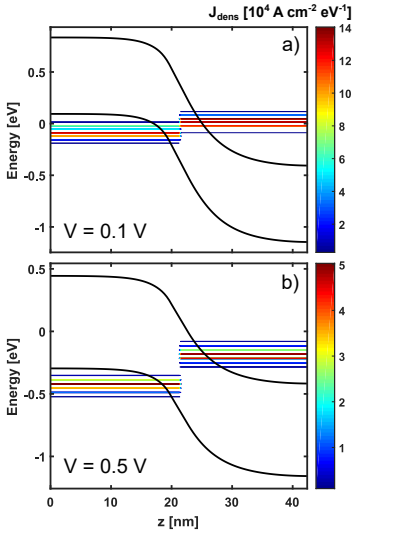


Figure 6: Spectral current density at (a) V=0.1 V and (b) V=0.5 V for the InGaAs Esaki diode of Fig. 5.

Acknowledgments. Work funded by the ANR project "GeSPAD" (ANR-20-CE24-0004).

On the Feasibility of DoS-Engineering for Achieving Sub-60 mV Subthreshold Slope in MOSFETs

J.M. Gonzalez-Medina[1], Z. Stanojevic[1], Z. Hou[2], Q. Zhang[2], W. Li[2], J. Xu[2] and M. Karner[1]
[1]Global TCAD Solutions, Vienna, email: jm.gonzalezmedina@globaltcad.com. [2]HiSilicon Technologies, Shenzhen

***Abstract*—We present the operating principle of an ideal Cold Source Field Effect Transistor and check the DoS source engineering impact on its subthreshold slope. The Subband Boltzmann Transport Equation is solved and the resulting transfer curves in the ballistic regime are presented, as well as those including the effects of scattering. The inclusion of scattering reveals its importance in the rethermalization of the cold carriers at the source extension and the degradation in the static leakage of the device. Finally, we show the impact in the SS when substituting the semiconducting source extension by the cold metal.**

I. Introduction

In the scaling process of the field effect transistors (FET), several challenges are faced. One of them that still persists is achieving a sub-60 mV/decade subthreshold swing (SS) while preserving a high enough on-off current ratio. There are several designs, such as the Tunnel FET (TFET), that can achieve a good SS, but with an on-off current ratio penalty due to the band-to-band tunneling process required to turn on the device. Other approach consists of engineering the density of states (DoS) by *filtering* the distribution function (DF) at the source contact. The idea relies on reducing the DoS at the source for high energies, in order to cut-off the Boltzmann tail of the DF. This can be achieved through several process, like (i) band to band tunneling in Cold Source FET (CSFET) [1, 2], (ii) mini-bands in a Superlattice (SLFET) [3], (iii) or a feature of the contact band structure, like the Dirac Source (DSFET) [4–6], among others. In this work, we study and simulate an ideal CSFET device, check the efectiveness of this DoS engineering and compare it to a Schottky Barrier CSFET (SBCSFET).

II. CSFET Results

The device (Fig. 1 (a)) consists of an idealized Silicon nanowire CSFET transistor with ohmic contacts, without additional tunneling processes or other contact resistances. At the source, we set a null DF/DoS above $E_{\text{cut-off}}$, which in this case is 0.1 eV above the source Fermi energy. In this example, below this $E_{\text{cut-off}}$ energy, the transmission probability is 1. The working process is shown in Fig. 2 (a) and (b): at on state, the barrier at the channel is low enough, so carriers below $E_{\text{cut-off}}$ can freely move. At off state, the barrier at the channel prevents the carrier flow, and this process is accelerated by the narrower window provided by the engineered DoS at the source. With a high enough barrier, ballistic current can be blocked, apart from Source-Drain tunneling. The device is simulated using the subband Boltzmann Transport Equation (SBTE) implemented in our Nano Device Simulator (NDS) [7].

As shown in Fig. 1 (b), the IV curves reflect how the ballistic current has a steep SS in the off state compared to a regular MOSFET (down to 17 mV/decade in this ideal case). The narrower DoS also reduces the on current with respect to the same device without DF filtering, both at low and high bias conditions. This simulation shows how, in an ideal ballistic situation, this concept can fit with the theoretical behaviour of a desirable sub-60 mV/decade. The blocking process can be seen in the distribution function at the energy-space grid depicted in Fig. 2 (b), where it is clear that the DoS blocking bans the carriers from flowing above the barrier.

Next, the scattering process is included, thus carriers can heat up and have energies above $E_{\text{cut-off}}$, resulting in a flow of carriers above the channel barrier, as can be seen in Fig. 2 (c), and in the current density plot in (a). This process takes less than 1 ps in Silicon [8]. The impact of this in the SS is shown in Fig. 3, where it remains clear that this DoS filtering can have a practically null impact in the device performance with respect to a regular MOSFET in steady state conditions.

III. Schottky Contact Results

One way to avoid the rethermalization of the carriers at the source extension is substituting this semiconducting section by the cold metal (Fig. 1 (c)), resulting in a Schottky contact (see energy plots in Fig. 3 (d) to (g)). The tunneling process is evaluated using the WKB method explained in [7]. As now the current is mostly tunnel current, and as it mostly depends on the barrier thickness and very weakly on the temperature, the SS is mostly restored, but with the expected penalty in the on current due to the extra resistance added by the tunneling.

IV. Conclusions

An ideal CSFET implemented in a Silicon nanowire is studied. In presence of scattering, the sub-60 mV slope is only transient and would recover to over 60 mV in a few ps. Steep-slope devices are only useful if they can maintain a sub-60 mV slope over long periods of time, i.e. between switching events, in order to suppress static leakage. Steady-state steep slope only appears in ballistic simulation, but scattering inclusion severely changes the result. Approaches based on DF filtering or DoS engineering of the source in an otherwise regular MOSFET device (CSFET, SLFET, DSFET, etc.) cannot meet this requirement. Alternative SBCSFET restore the steep slope because it becomes temperature independent, but brings back the on current penalty present in tunnel devices.

References

[1] E. G. Marin *et al.*, ACS Nano **14**, 1982 (2020).
[2] W. Gan *et al.*, IEEE Transactions on Electron Devices **67**, 2243 (2020).
[3] P. Maiorano *et al.*, *ESSDERC* (IEEE, 2013).
[4] C. Qiu *et al.*, Science **361**, 387 (2018).
[5] F. Liu *et al.*, IEEE Transactions on Electron Devices **65**, 2736 (2018).
[6] Z. Tang *et al.*, Nano Letters **21**, 1758 (2021).
[7] Z. Stanojevic *et al.*, IEEE Transactions on Electron Devices **68**, 5400 (2021).
[8] M. Lundstrom, *Fundamental of Carrier Transport* (Cambridge University Press, New York. USA, 2009).

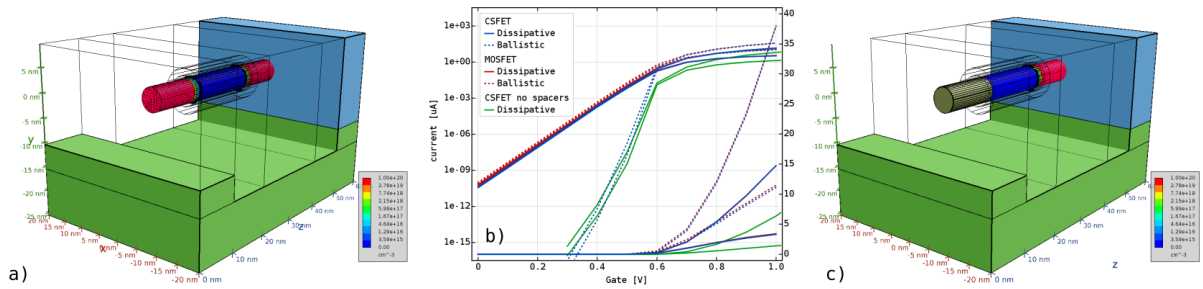


Fig. 1: (a) Simulated Silicon nanowire transistor device that acts as a CSFET. (b) Resulting transfer curves in (solid) dissipative and (dotted) ballistic regime. The figure also includes the IV curves in dissipative regime for a CSFET with source extension substituted by the cold metal, shown in (c).

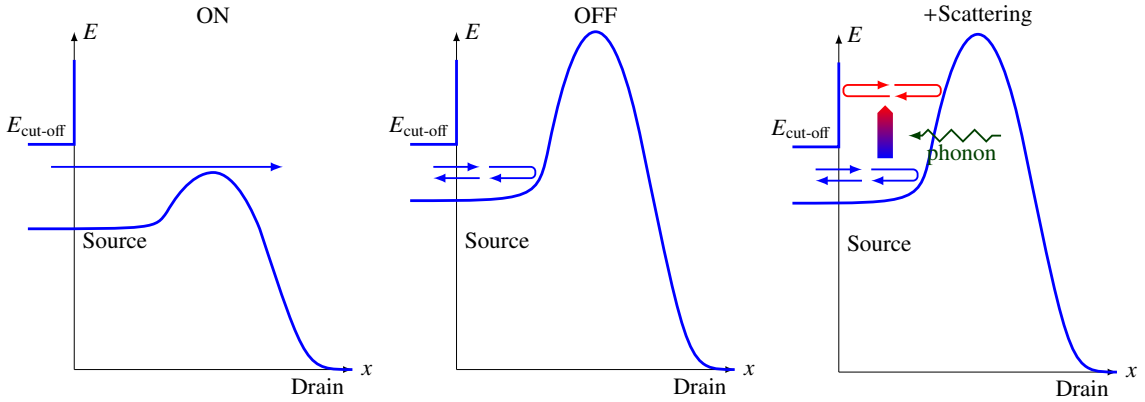


Fig. 2: Working principle of an ideal Cold Source FET device (a) ON state, (b) OFF state in ballistic situation. The inclusion of scattering effect (c) can potentially increase the energy of the carriers at the source.

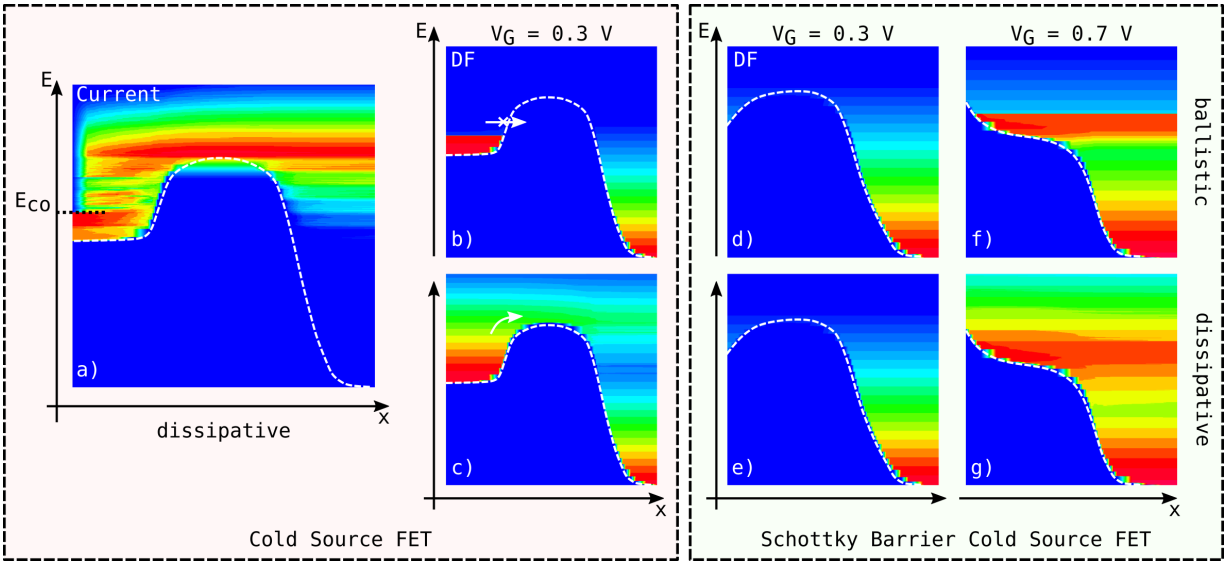


Fig. 3: (a) Current density of a CSFET in energy-space plot at V_G = 0.3 V. Distribution function in energy-space for V_G = 0.3 V in ballistic (b) and dissipative (c) regimes. The DoS cut-off prevents the carrier at higher energies to flow above the channel barrier only in ballistic regime. Scattering rapidly rises their energy, allowing them to flow. (d) to (g) Similar distribution function plots in a device with a source Schottky contact. The device turns to behave essentially as a Tunnel FET with an extra energy cut-off.

Theoretically probing the relationship between barrier length and resistance in Al/AlOx/Al tunnel Junctions.

Paul Lapham, Vihar Georgiev
Device Modelling Group
James Watt School of Engineering
University of Glasgow, Glasgow G12 8QQ, United Kingdom
E-mail: vihar.georgiev@glasgow.ac.uk

INTRODUCTION

One of the most popular qubit architectures is the superconducting qubit, which relies on the physics of the Josephson Junction (JJ) [1]. The main challenge faced in quantum computing is decoherence, often linked to two level defects within the JJ structure [2]. Josephson Junctions are tri-layer systems that consist of two superconductors separated by a thin insulating barrier, typically Al/AlO_x/Al. The thin oxide barrier is amorphous which is considered one of the main causes for noise and decoherence in qubit circuits [3]. There is still a poor understanding from an atomic perspective of how the structure of the amorphous oxide affects performance and subsequent failures in qubit applications.

There are many different variables within JJs structure that can sensitively influence performance [4]. Barrier Length is considered a key junction parameter, with typical barrier lengths ranging between 1-2 nm. However, the physical barrier length is difficult to determine and control. It has been shown that the effective barrier length can be orders of magnitude smaller than the physical barrier length [5]. Thus, it is critical to understand the nature of transport through the barrier. In this work we combine Tight Binding Density Functional methods (DFTB) and Non-Equilibrium Greens Function (NEGF) for an efficient method to probe theoretically the influence of the barrier length and atomic structure on the transport of Josephson Junctions.

The aim is to understand from an atomistic level how transport is affected by the atomic structure of the amorphous barrier. Through our simulation, we can explore the existence of an "effective barrier length" over a physical one and understand what structural effects lead to the reduction of this parameter.

SIMULATION METHODOLOGY

All calculations were carried out using the QuantumATK-2021 software.[6] The electronic properties of the Josephson Junctions were simulated with DFTB using the "matsci-0-3" parameter set [7], [8]. Although a semi-empirical method parameterized for certain systems with DFT, it has been shown to be transferrable and suggests it can describe the important physics of the systems with reasonable accuracy whilst significantly reducing the computational cost [9]. The transport properties of the junction were studied using non-equilibrium Green's function (NEGF), a powerful tool for a three-dimensional, atomistic treatment of transport. The amorphous oxide barrier was built by simulated annealing method using Molecular Dynamics as described several times in the literature using the "ReaxFF" force field [10].

RESULTS AND DISCUSSION

Fig. 1 shows an example of Al/AlOx/Al tunnel junctions studied in this work. The amorphous barrier is generated through molecular dynamics method before being geometrically relaxed between bulk Al electrodes using DFTB methods. The final device is studied by combining DFTB and NEGF to understand the electron transport in the device.

Fig. 2 shows the computed resistance for the junctions against the different barrier lengths. Data from 13 junctions are plotted. The relationship between physical barrier length and resistance is not straightforward. There is a slight exponential relationship, as guided by the red-dashed line. However, there is considerable variability in resistance values for junctions with similar physical barrier lengths, suggesting different "effective barrier lengths". For example, two junctions with 12 Angstrom barrier length show four times difference in the resistance.

Fig. 3 compares the computed transmission spectra for two junctions (A and B) with similar barrier lengths (12 and 12.3 Angstrom) but considerably different resistance values (70.8 and 16.5 kΩ). Despite similar barrier length and density, the inherent differences in the disorder within the barrier results in significantly different transmission spectra from which the Resistance value is calculated.

Fig. 4 shows the computed current-voltage (I-V) curves for Junction A and B between an applied bias of 0-1V. The I-V characteristics are consistent with the respective resistances, with the current for Junction B considerably larger than that for Junction A. Highlighting the sensitivity of transport to the atomic structure of the amorphous barrier.

Fig. 5 illustrates the Projected Local Density of States (PLDOS) for the two junctions. Consistent with the differing Resistances of the two junctions, Junction B shows small gaps ('pinholes') in DOS around the Fermi compared to Junction A. These differences illustrate the strong influence of local atomic structure on the transport properties of the junctions.

Fig. 6 shows the transmission pathways at the Fermi level for both Junction A and Junction B. Analysis of transmission pathways is a key advantage of our atomistic approach. It illustrates, the dominant atoms an area of the amorphous barrier responsible for transport. Here, an insight can be gained on why the resistance varies considerably despite similar barrier length, stoichiometry, density. Metallic hotspots are evident, consistent with previous work in the literature.

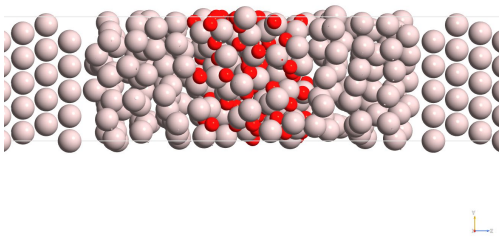


Figure 1. An example of one of the Al/AlOx/Al Junctions studied. Grey-Aluminium, Red- Oxygen.

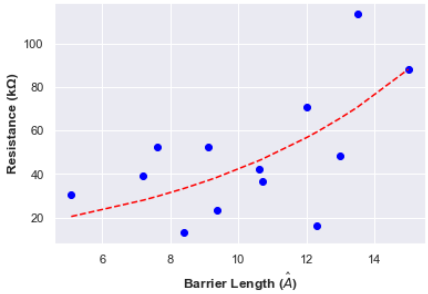


Figure 2. Computed Resistance vs Barrier Length. The red line offers a crude fit to guide the eye to the trend.

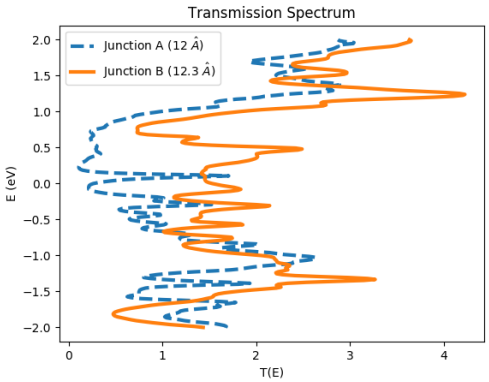


Figure 3. Computed Transmission Spectra for two Junctions with similar barrier length.

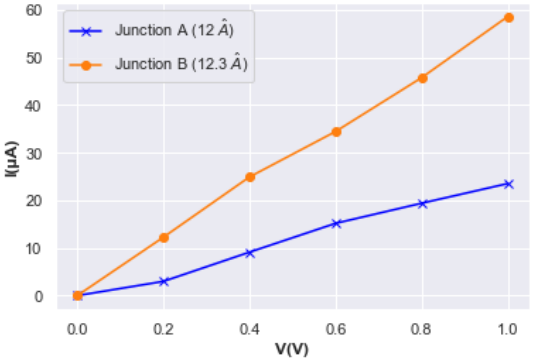


Figure 4. Computed I-V Curves for Junction A and Junction B which differ in barrier length by only 0.3 Angstroms.

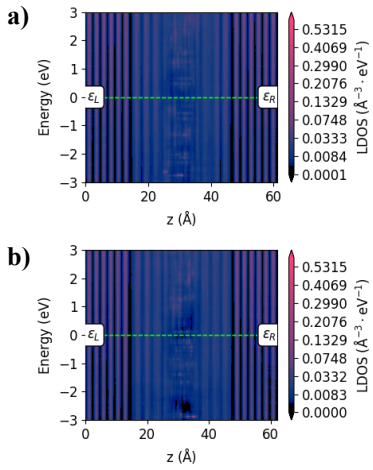


Figure 5. Projected Local Density of States (PLDOS) at zero bias for a) Junction A b) Junction B.

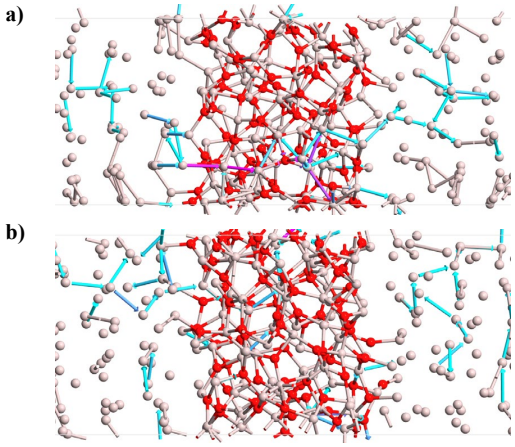


Figure 6. a) Transmission Pathways at Fermi level for Junction A b) Transmission Pathways at Fermi level for B, illustrating considerable differences in tunnelling despite similar barrier length.

REFERENCES

[1] M. Kjaergaard *et al.*, “Superconducting Qubits: Current State of Play,” *Annu. Rev. Condens. Matter Phys.*, vol. 11, pp. 369–395, 2020, doi: 10.1146/annurev-conmatphys-031119-050605.

[2] C. Müller, J. H. Cole, and J. Lisenfeld, “Towards understanding two-level-systems in amorphous solids: Insights from quantum circuits,” *Reports Prog. Phys.*, vol. 82, no. 12, 2019, doi: 10.1088/1361-6633/ab3a7e.

[3] L. Zeng, D. T. Tran, C. W. Tai, G. Svensson, and E. Olsson, “Atomic structure and oxygen deficiency of the ultrathin aluminium oxide barrier in Al/AlOx/Al Josephson junctions,” *Sci. Rep.*, vol. 6, no. March, pp. 1–8, 2016, doi: 10.1038/srep29679.

[4] C. E. Kim, K. G. Ray, and V. Lordi, “A density-functional theory study of the Al/AlOx/Al tunnel junction,” *J. Appl. Phys.*, vol. 128, no. 15, 2020, doi: 10.1063/5.0020292.

[5] L. S. Dorneles, D. M. Schaefer, M. Carara, and L. F. Schelp, “The use of Simmons’ equation to quantify the insulating barrier parameters in Al/AlOx/Al tunnel junctions,” *Appl. Phys. Lett.*, vol. 82, no. 17, pp. 2832–2834, 2003, doi: 10.1063/1.1569986.

[6] S. Smidstrup *et al.*, “QuantumATK: An integrated platform of electronic and atomic-scale modelling tools,” *J. Phys. Condens. Matter*, vol. 32, no. 1, 2020, doi: 10.1088/1361-648X/ab4007.

[7] J. Frenzel, A. F. Oliveira, H. A. Duarte, T. Heine, and G. Seifert, “Structural and electronic properties of bulk gibbsite and gibbsite surfaces,” *Zeitschrift fur Anorg. und Allg. Chemie*, vol. 631, no. 6–7, pp. 1267–1271, 2005, doi: 10.1002/zaac.200500051.

[8] L. Guimarães, A. N. Enyashin, J. Frenzel, T. Heine, H. A. Duarte, and G. Seifert, “Imogolite Nanotubes: Stability, Electronic, and Mechanical Properties,” vol. 1, no. 4, 2007.

[9] F. Spiegelman *et al.*, “Density-functional tight-binding: basic concepts and applications to molecules and clusters,” *Adv. Phys. X*, vol. 5, no. 1, 2020, doi: 10.1080/23746149.2019.1710252.

[10] M. J. Cyster, J. S. Smith, J. A. Vaitkus, N. Vogt, S. P. Russo, and J. H. Cole, “The effect of atomic structure on the electrical response of aluminium oxide tunnel junctions,” *arXiv*, vol. 013110, pp. 1–10, 2019, doi: 10.1103/physrevresearch.2.013110.

3D Feature-Scale Modeling of Highly Selective Fluorocarbon Plasma Etching

Frâncio Rodrigues*, Luiz Felipe Aguinsky*, Andreas Hössinger†, and Josef Weinbub*
*Christian Doppler Laboratory for High Performance TCAD at the
Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, 1040 Wien, Austria
†Silvaco Europe Ltd., Compass Point, St Ives, Cambridge, PE27 5JL, United Kingdom
Email: rodrigues@iue.tuwien.ac.at

***Abstract*—Modern semiconductor fabrication technology requires fluorocarbon dry etching of high aspect ratio silica-based structures with stringent material selectivity. Dry etching development faces challenges due to the lack of understanding of surface mechanisms and the etchant flux distribution on the feature-scale. We present a three-dimensional, TCAD-compatible, feature-scale modeling methodology to study these effects. We apply our methodology to a highly selective and high aspect ratio etching of SiO_2 by CF_4/C_4F_8 using Ru metallic mask layers. We are able to accurately reproduce the etch rates, topography, and critical dimensions of the experiment. We show that our methodology can precisely generate three-dimensional structures and can be used to prototype and study novel etching processes.**

Introduction Plasma etching is one of the key enabling fabrication techniques that is challenged to produce ever-smaller critical dimensions (CDs) with demanding high aspect ratios (HARs) and high selectivity while etching different materials [1]. The thus manifesting complexity of plasma etching often requires long process development cycles for new technologies, demanding continued modeling progress to enable further process optimizations. Feature-scale modeling of plasma etching is a powerful tool to investigate surface phenomena and etch rate topography dependencies. In particular, fluorocarbon plasma etching is notably challenging to model due to the simultaneous etching and polymer deposition mechanisms. Nonetheless, the deposition of a protective polymer layer is what allows the fabrication of HAR structures, stressing the importance of accurately modeling the interplay between the etchant and protective reactants. Another challenging aspect is visibility effects due to the distribution of incoming reactants, which cause unwanted aspect ratio dependent etching (ARDE) [1]. Therefore, controlling phenomena such as ARDE and improving the selectivity over mask materials in HAR vertical dry etch processes remains a challenge. Surface reactions and topography dependencies are thus a major research focus by modelling groups [2]–[4]. In particular, experimentalists are exploring different materials that can be used as thinner and inert masks to improve etch selectivity [5].

In this work, we study one of these novel stacks of materials with our feature-scale modeling methodology by simulating a three-dimensional (3D) SiO_2 via with a Ru mask etched by CF_4/C_4F_8 in an inductively coupled plasma (ICP) reactor [5]. We calibrate our models based on experimental data, characterize the etch rate of the materials, reproduce the experimental CDs, and show how our TCAD-compatible methodology can be used to accurately prototype processes with novel materials into 3D structures that can be integrated into TCAD process/device simulations.

Methodology We apply our previously devised surface reaction model [6] that is implemented into Silvaco's *Victory Process* simulator [7], use the tool's ray-tracer [8] for flux calculations, and its level-set engine [9] for the topography evolution (Fig. 1). The ray-tracer calculates the flux of incoming reactants taking into account the source angular distribution and shadowing effects. The resultant fluxes serve as an input to the surface reaction model, described by a Langmuir set of coverage equations (1-3). The calculated polymer coverage value (4) determines if a polymer deposition (5) or a surface etching (6) occurs. The level-set engine is responsible for applying the rates to the surface and updating the feature geometry.

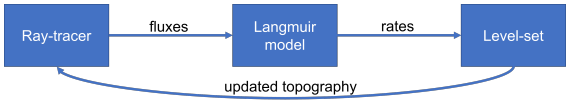


Fig. 1: Feature-scale modeling methodology: The ray-tracer evaluates the local fluxes. Langmuir equations use these fluxes to calculate the etch or deposition rates and the level-set engine evolves the surface accordingly.

The complex mixture of reactants generated by the plasma is abstracted into three functional species: neutrals (n), ions (i), and polymers (p). The subscript n/p indicates the coverage of neutrals on a polymer substrate. Neutrals represent the etchants, polymers the etching inhibitors and the depositing species of a polymer layer, and ions are responsible for the sputtering and reactive ion etching (RIE) mechanisms. The substrate material coverage by n and p is described by a Langmuir adsorption model (1-2). The polymer material has its own independent Langmuir equation because it can also be etched or deposited (3). Each term of the equations represents a mechanism of adsorption or etching (RIE, sputtering, evaporation). The coverages $\Theta_{n,p}$ are defined as the fraction of substrate surface sites with adsorbed n or p species. $\Theta_{n/p}$ describes the coverage of polymer by neutral species to take the polymer etching into account. A steady-state approximation ($\frac{d\Theta_{n,p,n/p}}{dt} = 0$) is used to solve equations (1-3) because the surface evolves very slowly compared to the time it takes for the adsorption and desorption mechanisms to take place [2].

The other quantities present in (1-6) are the fluxes $J_{n,i,p}$, the sticking coefficients $S_{n,p,n/p}$, the stoichiometric coefficient k_n, the substrate densities for polymer (ρ_p) and SiO_2 (ρ_{SiO_2}), the rates R_{dep} and R_{etch}, the sputtering yield Y_s, and the RIE yields $Y_{n,n/p}$. J_{ev}, the evaporation flux, and k_{ev} its stoichiometric constant, are used to model the thermal evaporation etching mechanism [2].

$$\frac{d\Theta_n}{dt} = J_n S_n (1 - \Theta_n - \Theta_p) - J_i Y_n k_n \Theta_n - J_{ev} k_{ev} \Theta_n \quad (1)$$

$$\frac{d\Theta_p}{dt} = J_p S_p - J_i Y_{n/p} \Theta_p \Theta_{n/p} \quad (2)$$

$$\frac{d\Theta_{n/p}}{dt} = J_n S_{n/p} (1 - \Theta_{n/p}) - J_i Y_{n/p} \Theta_{n/p} \quad (3)$$

$$\Theta_p = \frac{J_p S_p}{J_i Y_{n/p} \Theta_{n/p}} \quad (4)$$

$$R_{dep} = \frac{J_i Y_{n/p} \Theta_{n/p} - J_p S_p}{\rho_p} \quad (5)$$

$$R_{etch} = \frac{1}{\rho_{SiO_2}} (J_i Y_n k_n \Theta_n + J_i Y_s (1 - \Theta_n - \Theta_p) + J_{ev} k_{ev} \Theta_n) \quad (6)$$

Θ_p (4) defines whether a deposition or etch step occurs. If $\Theta_p \geq 1$ the surface is completely covered by polymers and a deposition occurs with a rate of (5). However, if $\Theta_p < 1$ the substrate is etched at a rate given by (6).

Results We use the experimental data from [5] to calibrate and validate our simulation results. The experimental setup of the ICP reactor is defined in Tab. I and the initial feature's two-dimensional (2D) cross-section in Fig. 2(a). A 3D simulation was performed and, to improve performance, we explored the trench symmetry and ran the simulation only on a half 3D trench and mirrored the result to represent the full trench. To estimate the neutrals and polymer flux, we adapted values from the steady-state densities of neutral species reported by [10] to the setup shown in Tab. I, resulting in $J_n = 1 \times 10^{17}\, cm^{-2} s^{-1}$ and $J_p/J_n = 0.14$. This J_p to J_n ratio indicates a strong polymerization regime, where $\mathrm{F/C} < 3$, which is expected for anisotropic etching applications [10]. The remaining stoichiometric constants, J_e, and the sputtering functions and parameters for SiO_2 use values from our previous work [6]. Therefore, the fitting parameters are J_i and the angular distributions of the $J_{n,p,i}$ sources.

For the neutral and polymer particles we use a constant flux value across the surface and for the ions a sharp *von Mises* source angular distribution with a shape parameter of 250. The neutrals' and polymers' constant approximation is valid for our case of particles with low sticking values ($S_{n,p,n/p} = 0.1$) impinging onto HAR structures [11]. The J_i value of $1.4 \times 10^{16}\, cm^{-2} s^{-1}$ was found by fitting the etch rate and the shape of the final trench to the experimental data.

We ran the simulation for the total etch time of $94\,s$ and achieved the profile in Fig. 2 (b). The etch rate for SiO_2 ($332\,nm/min$) and the selectivity of SiO_2/Ru (78) are within $3\,\%$ of the reported values of $324\,nm/min$ and 72.5, respectively [5]. The simulated trench CDs were also in excellent agreement: The depth ($520\,nm$), width at the bottom ($58\,nm$), and width at the half-height point ($84\,nm$) are all within $5\,\%$ from the experimental results [5].

In Fig. 3, the 3D half-trench was mirrored several times to build an array of trenches. It shows the capability of our methodology to reproduce 3D structures based on physical simulations which can serve as an input for larger TCAD process/device simulations. In Fig. 3 the expected symmetry with regards to the trench length is also evident, showing that our flux and topography models can be reliably used in 3D structures.

TABLE I: ICP etch setup with a C_4F_8/CF_4 mixture [5]

Parameter	Value
Total gas flow	$60\,SCCM$
C_4F_8 content	$12.5\,\%$
Substrate power	$75\,W$
Pressure	$5\,mTorr$
Time (t)	$94\,s$

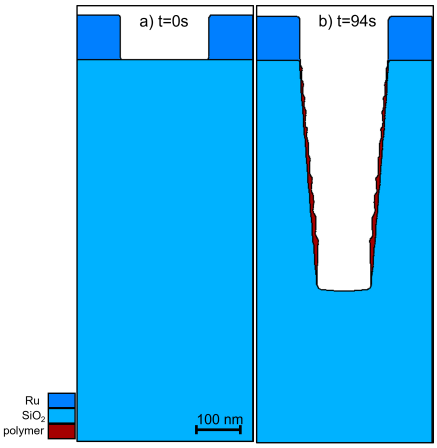


Fig. 2: a) 2D cross-section of the initial feature shape with a $400\,nm$ opening and a Ru mask height of $100\,nm$. b) 2D cross-section after total etch time of $94\,s$ with the protective polymer layer that is necessary for anisotropic structures.

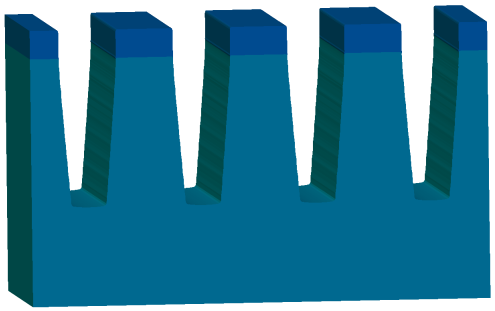


Fig. 3: A 3D array of trenches was built by mirroring our final profile and by stripping the polymer layer. Our methodology can be used to generate full 3D structures based on physical simulations for subsequent TCAD process/device simulations.

I. Acknowledgment

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged.

References

[1] T. Iwase, Y. Kamaji, S. Y. Kang *et al.*, "Progress and perspectives in dry processes for nanoscale feature fabrication: Fine pattern transfer and high-aspect-ratio feature formation," *Japanese Journal of Applied Physics*, vol. 58, no. SE, p. SE0802, 2019.

[2] A. L. Magna and G. Garozzo, "Factors affecting profile evolution in plasma etching of SiO_2: Modeling and experimental verification," *Journal of The Electrochemical Society*, vol. 150, no. 10, p. F178, 2003.

[3] S. Huang, C. Huard, S. Shim, S. K. Nam, I.-C. Song, S. Lu, and M. J. Kushner, "Plasma etching of high aspect ratio features in SiO_2 using $Ar/C_4F_8/O_2$ mixtures: A computational investigation," *Journal of Vacuum Science & Technology A*, vol. 37, no. 3, p. 031304, 2019.

[4] H. S. You, Y. G. Yook, W. S. Chang *et al.*, "Universal surface reaction model of plasma oxide etching," *Journal of Physics D: Applied Physics*, vol. 53, no. 38, p. 10, 2020.

[5] W. J. Mitchell, B. J. Thibeault, D. D. John, and T. E. Reynolds, "Highly selective and vertical etch of silicon dioxide using ruthenium films as an etch mask," *Journal of Vacuum Science & Technology A*, vol. 39, no. 4, p. 043204, 2021.

[6] F. Rodrigues, L. F. Aguinsky, A. Toifl, A. Scharinger, A. Hössinger, and J. Weinbub, "Surface reaction and topography modeling of fluorocarbon plasma etching," in *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2021, pp. 229–232.

[7] Silvaco, "Silvaco Victory Process," 2022. [Online]. Available: www.silvaco.com/tcad/victory-process-3d/

[8] P. Manstetten, J. Weinbub, A. Hössinger, and S. Selberherr, "Using temporary explicit meshes for direct flux calculation on implicit surfaces," *Procedia Computer Science*, vol. 108, pp. 245–254, 2017.

[9] J. A. Sethian, *Level set methods and fast marching methods*. Cambridge University Press, 1999, vol. 3.

[10] N. Lim, A. Efremov, and K. Kwon, "A comparison of CF_4, CHF_3 and $C_4F_8 + Ar/O_2$ inductively coupled plasmas for dry etching applications," *Plasma Chemistry and Plasma Processing*, vol. 41, no. 6, pp. 1671–1689, 2021.

[11] A. Yanguas-Gil, *Growth and transport in nanostructured materials*. Springer, 2017.

DTCO Flow for Air Spacer Generation and its Impact on Power and Performance at N7

L. Filipovic[†*], O. Baumgartner[‡], J. Piso[†], J. Bobinac[†], T. Reiter[†], G. Strof[‡], G. Rzepa[‡], Z. Stanojevic[‡], M. Karner[‡]
[†]Institute for Microelectronics, TU Wien, 1040 Vienna, Austria
[‡]Global TCAD Solutions GmbH., 1010 Vienna, Austria
[*]Email: filipovic@iue.tuwien.ac.at

I. Introduction

The integration of process simulations with device and circuit simulations through design-technology co-optimization (DTCO) is essential for the successful design of future semiconductor devices and technologies [1]. In this manuscript, we describe a novel DTCO flow and apply it to study the impact of spacers with an air gap (AG), known as air spacers (ASs) on the circuit-level power and performance at the 7 nm node, using a 5-stage ring oscillator (RO) as an example circuit. The effective spacer capacitance, C_{eff}, is a limiting factor in the achievable frequency $f \propto 1/C_{\text{eff}}$ and power $P \propto C_{\text{eff}}$ of the oscillator [2].

It is essential to understand process variations and the limitations in reducing the capacitance imposed by the fabrication of the AS, when designing the RO circuit. The main fabrication parameters which impact the AG geometry are the sticking coefficient s and the thickness of the conformally-deposited SiN layer t_c which determines the width of the trench prior to non-conformal deposition. Since s is directly related to the fabrication condition in the chemical vapor deposition (CVD) chamber, such as pressure and temperature [3], this provides a direct link between circuit-level performance and the fabrication conditions.

II. Simulation Flow

The simulation of the fabrication-induced variation in the RO performance using physics-based models is not feasible, since physical process models require a time-intensive Monte Carlo (MC) ray tracing and level set (LS) approach. Therefore, we first perform a set of physical simulations in order to generate an analytical model for non-conformal CVD [4] which is based on these physics-based models, as implemented in ViennaPS [5]. This analytical model is subsequently applied in the full DTCO flow in order to study the impact of the fabrication parameters on the circuit performance. The critical steps in the workflow, as shown in Fig. 1, are described in this section.

A. Physical Topography Simulation

The physical simulation for the generation of the AG in the spacer is based on a LS powered topography simulator, together with top-down MC ray tracing for the simulation of non-conformal CVD [6]. Initially, a conformal layer of width t_c is deposited in the spacer trench, which can be modeled using physical or analytical approaches. After this, a non-conformal CVD is simulated using a single-particle approach, where the particle has a specific sticking coefficient s which describes its propensity to adsorb onto the surface [7]. Higher s values represent higher non-linearities which ensure that the gap is pinched off at the top. Several physical simulations are performed while varying t_c and s in order to subsequently devise a fast analytical model using the generated geometries.

B. Geometrical Description of the Air Gap

The principal aim of the analytical model is to reproduce the geometrical shape of the AG inside the spacer by reproducing the pinch-off height (POh), bottom height (Bh), and air gap width (AGw) from the physical CVD model (Fig. 2). This model also applies a linear interpolation for the air gap's geometrical parameters for s and t_c values which have not been simulated with the physical model. The air gap's shape is represented using a superellipse centered at $(0, 0)$ with radii r_x and r_y along the x and y axes, respectively, using the equation

$$y = \pm r_y \sqrt[4]{1 - |x/r_x|^4}. \tag{1}$$

C. Analytical Topography Simulation

In our DTCO flow, the complete AS is generated by first assuming a fully-filled SiN spacer and then performing a Boolean operation to remove the AG geometry from the spacer region. The AG geometry follows Eq. (1) while its vertical placement depends on the physically-modeled values of POh and Bh. This method allows to reproduce the physical AS model with high accuracy, as shown in Fig. 3, while requiring a fraction of the simulation time. The analytical model showed a speedup of about $100\times$, which is consistent with our previous studies [4].

D. Capacitance Extraction

The capacitance across the generated AS is calculated by solving the Poisson equation, which allows to extract a relationship between the capacitance and the chosen fabrication parameters. Ultimately, the calculated capacitance is used to extract an effective relative permittivity ε_{eff} of the AS, which contains an AG surrounded by SiN. The range of ε_{eff} we observe for the tested fabrication conditions is from about 4.2 to 5.7, as shown in Fig. 4, while a pure SiN spacer exhibits an ε_{eff} of 7.4.

E. Power-Performance Analysis

With the calculated ε_{eff} values, a power-performance analysis (PPA) of a 5-stage inverter RO logic cell is performed assuming varying fabrication conditions. The SPICE model cards are extracted automatically from the TCAD transistor characteristics and the parasitics network is calculated from the full three-dimensional (3D) RO cell using a field solver. Applying this method, any change in capacitance can be captured in a consistent manner [1], [8]. As expected, the results clearly show that the introduction of an air gap (AG) results in an improvement in the power and achievable frequency in all cases, as shown in Fig. 5.

III. Results and Discussion

We apply our developed DTCO framework on two fabrication flows, both of which are compatible with complementary semiconductor-metal-oxide (CMOS) technology. For one, the air spacer is formed prior to the deposition of middle of line (MOL) contacts (pre-MOL) [9] and for the second one the air spacer is formed after the deposition of MOL contacts (post-MOL) [2]. From Fig. 3 we note that both flows provide a similar air gap width, while the post-MOL-generated AG is shifted slightly up, due to the pinch-off location being slightly higher. A minimum conformal layer thickness of 2.5 nm was chosen because it was found that this allows for the pinch-off to stay within the spacer region. Otherwise, it may encroach into the inter-layer dielectric (ILD) layer between the MOL contacts.

The most significant impact on the spacer's effective relative permittivity is the AG width. This parameter is highly driven by the thickness of the conformally-deposited SiN. As can be observed from Fig. 4 the lowest permittivity is achieved when the sticking coefficient is high and the conformal layer thickness is low. These two factors essentially mean that the conformal deposition should set the width of the air gap, while the non-conformal deposition should only close the generated trench.

Finally, we observe the impact of the AS on a 5-stage inverter RO circuit using PPA, which is summarized in Fig. 5. We note that the inclusion of the air gap improves the performance by about $15\,\%$ in the worst case. With the presented framework, we are able to apply DTCO studies all the way from 3D physical process simulations through to geometric analysis of the spacer's topography and finally to thorough device and circuit analysis.

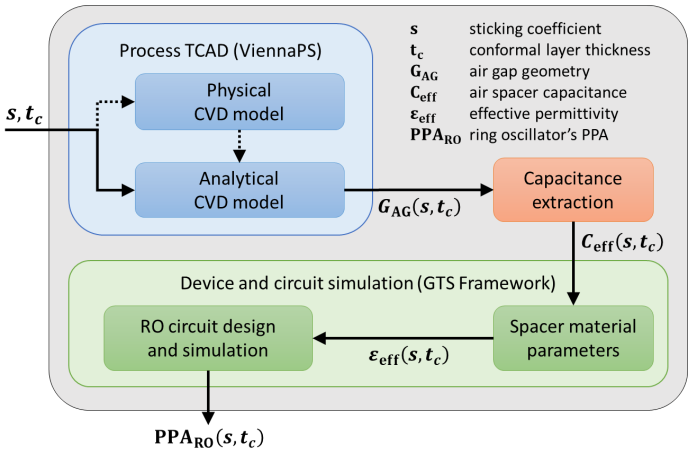


Fig. 1. Flowchart showing the presented DTCO workflow (solid arrow) for AS generation for a 5-stage inverter RO logic cell. The dotted arrow shows the development of the analytical model, which is based on a calibrated physical model. The main input parameters are the sticking coefficient s and conformal SiN layer thickness t_c, which are used to generate the air spacer geometry G_{AS}. The impact of the studied fabrication parameters (s, t_c) on the circuit power and performance are then provided using a PPA chart.

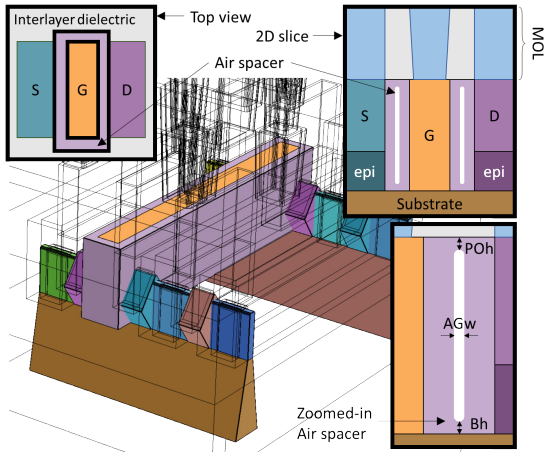


Fig. 2. Inverter cell with highlighted gate line and spacer regions. The insets show the spacer structure between source (S) and drain (D) regions, encircling the gate (G), below the MOL contacts. On the bottom right, the air spacer is shown with typical measurements for the pinch-off height (POh), bottom height (Bh) and the air gap width (AGw). These physical parameters are used to define the geometry of the air spacer G_{AS}.

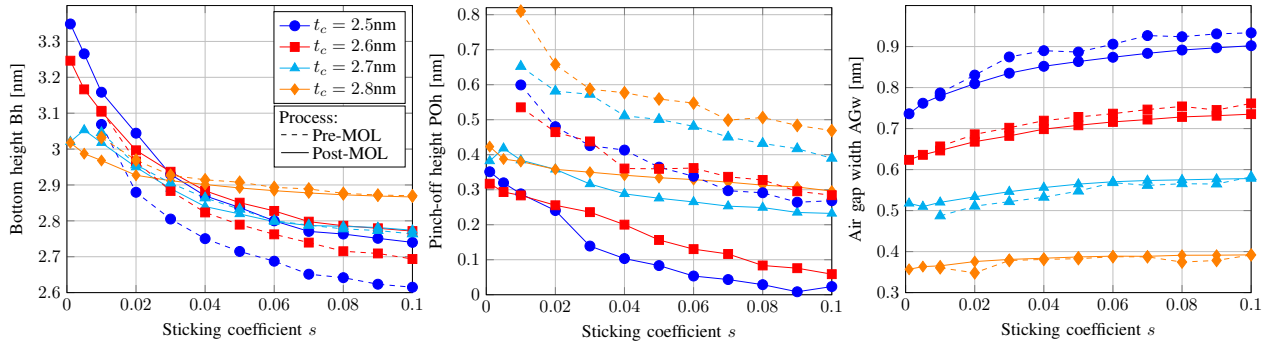


Fig. 3. Impact of s and t_c on the shape of the resulting air spacer geometry, including the bottom height (Bh), pinch-off height (POh), and the air gap width (AGw), as shown in Fig. 2. The symbols and lines show the results using the physical and analytical CVD models, respectively, while the dashed and solid lines show the results using the pre-MOL and post-MOL processes, respectively. We note that the analytical model is designed to replicate the physical model, while a linear interpolation is used to obtain the results between the physically-derived values. A clear impact of the two parameters is evident in all modeled scenarios.

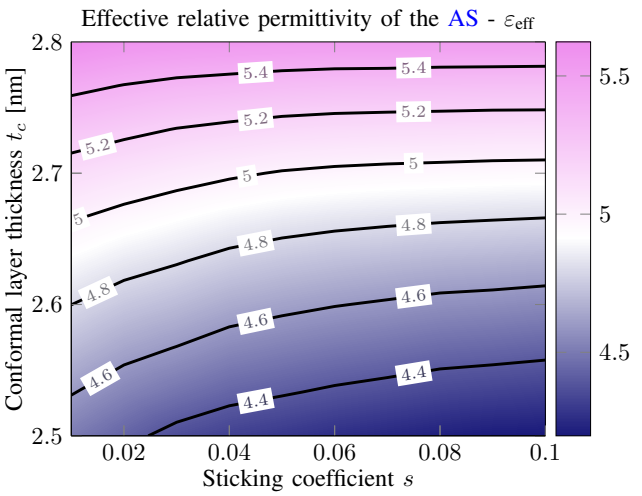


Fig. 4. The impact of the sticking coefficient s and conformal layer thickness t_c during air gap formation on the effective relative permittivity $\varepsilon_{\mathrm{eff}}$ of the air spacer using the post-MOL fabrication flow. We have also observed that the pre-MOL flow shows very similar results with $\varepsilon_{\mathrm{eff}}$ ranging from about 4.16 to 5.66 as s and t_c are varied. Therefore, increasing s and reducing t_c leads to lowest $\varepsilon_{\mathrm{eff}}$ values.

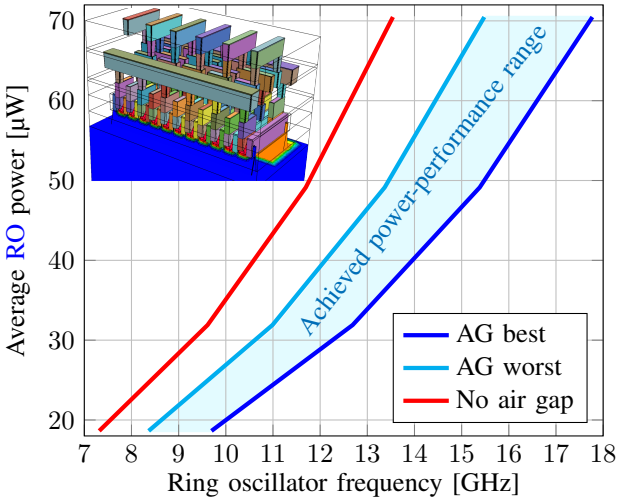


Fig. 5. Achieved power and performance for the RO with no air gap (AG) and with an AG under best and worst process conditions. The best condition corresponds to an $\varepsilon_{\mathrm{eff}}$ of 4.2 when $(s, t_c) = (0.1, 2.5\,\mathrm{nm})$. The worst condition has $\varepsilon_{\mathrm{eff}} = 5.7$ when $(s, t_c) = (0.02, 2.8\,\mathrm{nm})$ in the pre-MOL process. The case with no air gap corresponds to a completely filled SiN spacer. The inset shows the full RO cell.

[1] G. Rzepa *et al.*, in IRPS 2021. DOI:10.1109/irps46558.2021.9405172
[2] K. Cheng *et al.*, *IEEE Transactions on Electron Devices*, vol. 67, no. 12, pp. 5355–5361, 2020. DOI:10.1109/ted.2020.3031878
[3] N. Cheimarios *et al.*, *Archives of Computational Methods in Engineering*, vol. 28, no. 2, pp. 637–672, 2020. DOI:10.1007/s11831-019-09398-w
[4] L. Filipovic and X. Klemenschits, in SISPAD 2021. DOI:10.1109/sispad54002.2021.9592595
[5] X. Klemenschits *et al.*, "ViennaPS - Vienna process simulation library," [Online]. Available: https://github.com/ViennaTools
[6] X. Klemenschits, "Emulation and Simulation of Microelectronic Fabrication Processes," Doctoral dissertation, TU Wien, Vienna, Austria, 2022.
[7] T. S. Cale and G. B. Raupp, *Journal of Vacuum Science & Technology B*, vol. 8, no. 6, p. 1242, 1990. DOI:10.1116/1.584901
[8] "GTS Cell Designer," [Online]. Available: http://www.globaltcad.com/celldesigner
[9] K. Cheng *et al.*, in *IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 444–447. DOI:10.1109/iedm.2016.7838436

GPGPU MCII for high-energy implantation

Fumie Machida[1], Hiroo Koshimoto[1], Yasuyuki Kayama[1], Alexander Schmidt[2], Inkook Jang[2], Satoru Yamada[1], and Dae Sin Kim[2]
[1] *DS2 Lab, DS center, Samsung R&D Institute Japan (SRJ), Tsurumi-ku, Yokohama 230-0027, Japan*
[2] *CSE Team, Innovation Center, Samsung Electronics Co., Ltd., Hwasung-si, Gyeonggi-do 18448, Korea*

I. INTRODUCTION

The Binary-Collision-Approximation (BCA) based Monte Carlo (MC) simulation is widely used for predicting ion implantation profile. Various physical models have been proposed and improved by many researchers so far[1–3], and many efforts have been made to reduce the calculation time of the BCA based MC simulators[4,5]. Nevertheless, the calculation time is still an issue for high energy implantations, e.g. deep photodiode or deep well isolation formation, because of many collision events that have to be considered within BCA MC simulation. To resolve this problem, we have developed the GPU implementation of the BCA based MC simulator (GPU-MCII). In this paper, the GPU performance, its drawbacks and a technique reducing these drawbacks are shown.

II. METHODOLOGY

BCA MC simulation flow considers that energetic particles enter from the top of simulation domain, lose their energy due to successive collisions with the target material atoms nuclei and inelastic electronic energy losses until they come to rest. Thus a MC simulation cycle includes the following calculation steps:
(1) search target atoms
(2) calculate nuclear/electronic energy loss and direction of a projectile ion after collision based on BCA
(3) update next particle position
(4) calculate amount of energy transferred during collision and generated crystal damage (in crystalline materials only)
(5) check criterion of trajectory splitting[5]
(6) update dopant profile

This MC cycle is repeated until the energy of particle reaches the cut-off energy (5eV). Due to large number of real ions that are typically implanted in semiconductor technology (dose is in range $10^{11}\mathrm{cm}^{-2} \sim 10^{15}\mathrm{cm}^{-2}$), in MC simulation each MC particle corresponds to many real ions and so implanted dose and damage are scaled accordingly. It is critically important to track dynamic changes of lattice damage since they are affecting subsequent MC particle collision probabilities at step (1) of the loop[3]. In the CPU implementation of BCA algorithm (CPU-MCII), a loop of the MC cycles is executed sequentially for every MC particle. Even though the multicore CPU computes multiple MC particles in parallel, the impact on damage accumulation is limited because the number of parallel executions is small. However, it would be problematic with GPU due to massively parallelized execution (Sec. III).

Implementing the MC cycle onto GPU without ingenuity introduces the branch divergence[6] which is the performance degradation caused by branch instructions. To avoid the branch divergence, we decompose the MC cycle into multiple GPU kernels which contain no branch codes for the difference of material crystallinity, and define a GPU batch which is a tuple of the GPU kernels and the MC particles for a material crystallinity. A GPU batch is bound with a chunk of GPU cores. While some GPU batches run, the GPU-MCII flushes out the result of GPU batches completed, screens the MC particles still in flight and loads a next step onto GPU batches vacant (Fig. 1). In this way, the GPU-MCII carries out individual calculations for a mass of MC particles at once.

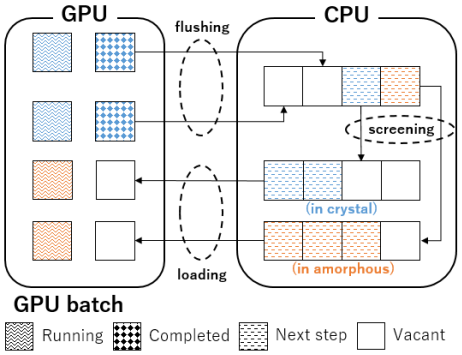


FIG. 1: Schematic of the GPU-MCII

III. SLOWCOACH SCHEME

In order to reduce the impact on damage accumulation as discussed in Sec. II, we introduced the slowcoach scheme with multiple slowcoach approaches (SA). SA is a method to use not full size of a GPU batch but the specified smaller size . It makes GPU utilization being lower, but can generate damage little by little. The keys to obtain accurate profile are the optimal task size and the end condition of the SA. To define them, we firstly take sample data with a small number of particles N_s using the SA (Fig. 2). Using maximum damage density $C_{d,max}$ from the sample data and the damage threshold ($C_{d,th}$) for amorphization, we could estimate that the number of particles required for the damage density to reach to the threshold is $N_s \frac{C_{d,th}}{C_{d,max}}$. And then, the number of completed tasks to end the SA is defined by $A \cdot N_s \frac{C_{d,th}}{C_{d,max}}$. In the same way, the task size to be used in the SA is defined by $B \cdot N_s \frac{C_{d,th}}{C_{d,max}}$. Where, A and B are user parameters,

and $A \geq 1$ and $B < 1$ are desirable to generate damage slowly.

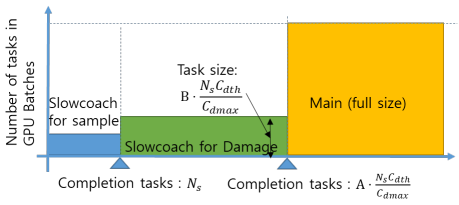


FIG. 2: Slowcoach approach schematic

Fig. 3 shows Arsenic (As) profiles obtained by CPU-MCII (calibrated to SIMS data) and GPU-MCII. Without SA, the tail profile is suppressed comparing to CPU-MCII because the peak area is amorphized in the early stage due to the concentration of damage generation. In reality, at early stage of implantation surface damage is small and some portion of ions have high chance of channeling, forming deep tail profile. When SA is used this effect is properly captured and the profile becomes similar to the result of CPU-MCII.

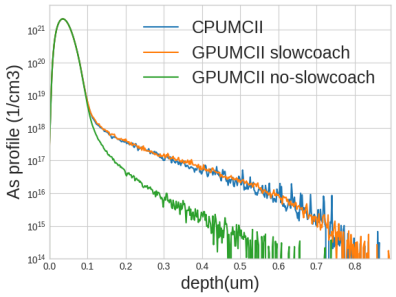


FIG. 3: Difference between with and without slowcoach (dose=8.0×10^{15}cm^{-2}, energy=50keV,A=4,B=0.1)

IV. BENCHMARK

We benchmarked deep implantation into a 3D CIS structure consisting of multiple materials using an unstructured mesh. The GPU-MCII and the CPU-MCII were performed with Nvidia Tesla V100 and 8 cores on Intel Xeon Gold 6146 3.2GHz. Fig. 4 shows As implantation results. Energy is a few MeV and dose is 5×10^{11}cm^{-2}. In this case high acceleration rate is expected and indeed simulation time of GPU-MCII is merely 3.8% of CPU-MCII for 1×10^{6} particles statistics and only 2.5% for 1×10^{8} particles due to reduced relative overhead for simulation initialization and finalization.

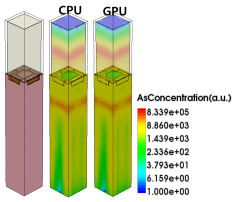


FIG. 4: deep implantation of As into 3D CIS

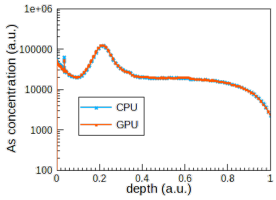


FIG. 5: 1D profile in depth direction

TABLE I: TAT comparison between CPU and GPU

Number of particles	calculation time[hour]		CPU×8/GPU
	CPU×8	GPU	
1×10^{6}	0.75	0.03	26.5
1×10^{8}	66.36	1.67	39.8

V. SUMMARY

We developed the GPU-MCII that utilizes the decomposition strategies to avoid the branch divergence performance degradation on GPU. Slowcoach Approach algorithm was implemented in order to mitigate the all-at-once damage accumulation overestimation due to massively parallel GPU execution. Notwithstanding the fact that parallel performance of GPGPU-MCII can degrade for low energies and high doses, when frequent CPU-GPU data sharing is needed to reset simulation conditions, our methodology is showing excellent results for high energy and low does implantation steps simulation, which are important for CIS technology optimization. For typical process conditions and modeling setup almost 40x simulation time acceleration can be achieved.

[1] J. F. Ziegler, J. P. Biersack, and M. D. Ziegler, *SRIM - The Stopping and Range of Ions in Matter*. James Ziegler, 2008.
[2] G. Hobler, H. Pötzl, L. Gong, and H. Ryssel, "Two-Dimensional Monte Carlo Simulation of Boron Impplantation in Crystalline Silicon," *Simulation of Semiconductor devices and process*, vol. 4, pp. 389–398, 1991.
[3] S. Tian, "Predictive Monte Carlo ion implantation simulator from sub-keV to above 10 MeV," *JAP*, vol. 93, pp. 5893–5904, 2003.
[4] Y. C. G. Wang, B. Odradovic, "A computationally efficient target search algorithm for a Monte Carlo ion implantation simulator," *Journal of Technology Computer Aided Design TCAD*, 1996.
[5] H. Kubotera, Y. Kayama, S. Nagura, Y. Usami, A. Schmidt, U. Kwon, K.-H. Lee, and Y. Park, "Efficient Monte Carlo Simulation of ion Implantation into 3D FinFET structure," 2014.
[6] W. W. Fung, I. Sham, G. Yuan, and T. M. Aamodt, "Dynamic Warp Formation and Scheduling for Efficient GPU Control Flow," in *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*, pp. 407–420, 2007.

Modeling Electrical Resistivity of CrSi Thin Films

K. Sonoda[1], N. Shiraishi[2], K. Maekawa[1], N. Ito[1], E. Hasegawa[2], and T. Ogata[1]
[1]Renesas Electronics Corporation, 751, Horiguchi, Hitachinaka, Ibaraki, 312-8511, Japan
[2]Renesas Electronics Corporation, 1-1-1, Yahata, Minami, Kumamoto, 861-4195, Japan
Email: kenichiro.sonoda.xc@renesas.com

***Abstract*—Electrical properties of CrSi thin films are modeled considering phase transitions and grain growth during annealing. The effective medium approximation is used to calculate the resistivity of the film which comprises several phases including grain boundaries. The phase transition from as-deposited amorphous to poly-crystalline including a meta-stable state leads to high resistivity within a certain range of annealing temperature.**

I. Introduction

CrSi based thin films are attracting attention because of their high resistivity and low temperature coefficient of resistivity (TCR). Electrical properties of CrSi thin films depend on microstructure of the constituting materials[1]. Post-annealing changes transport properties because of crystallization[2]. In order to optimize wafer process conditions to achieve desired transport properties, quantitative modeling of the effect of annealing on the electrical properties is required.

In this study, electrical properties of CrSi thin films are modeled considering phase transitions and grain growth during annealing. Experimental data with Si/Cr atomic ratio slightly larger than 2[1] are used for model validation.

II. Modeling

We formulate the phase transitions during annealing using temperature-dependent transition rates between phases. Volume fractions of each phase are obtained by solving rate equations. The resistivity is calculated based on the volume fractions.

The rate equations are derived from a set of phase transitions, as shown in Fig. 1, which is based on physical analysis results of the thin films annealed with various temperatures[1]. The initial phase is amorphous CrSi (a-CrSi). The final phase is polycrystalline $CrSi_2$ (p-$CrSi_2$), which is the stable phase of Cr-Si alloys with Si/Cr atomic ratio around 2[4]. The intermediate phase is polycrystalline Cr_5Si_3 (p-Cr_5Si_3). As the Si/Cr atomic ratio of the initial a-CrSi is slightly larger than 2, both p-Cr_5Si_3 and p-$CrSi_2$ regions are accompanied by polycrystalline Si according to the principle of mass-balance. The volume fractions of a-CrSi, p-Cr_5Si_3, p-$CrSi_2$, and Si are obtained by solving the rate equations.

The rate of Transition $i(=1,2,3)$ shown in Fig. 1 at annealing temperature $T_{\rm an}$ is expressed as $P_i = \nu((T_{\rm mp} - T_{\rm an})/T_{\rm mp})\exp(-E_{{\rm a}i}/(kT_{\rm an}))$ which is based on the expression for crystal growth[5] with the lattice vibration frequency $\nu = 1\times10^{13}\,{\rm Hz}$, the melting point $T_{\rm mp} = 1763\,{\rm K}$, and the activation energy $E_{{\rm a}i}$.

The physical analysis results also indicate that the higher the annealing temperature, the larger the grains of p-Cr_5Si_3 and p-$CrSi_2$[1]. The observed grain radius after annealing time $t_{\rm an}$ is expressed by $r_{\rm g} = k_{\rm gg0}\exp(-E_{\rm agg}/(kT_{\rm an}))t_{\rm an}^{1/2}$ with $k_{\rm gg0} = 1\times10^2\,{\rm nm/s^{1/2}}$ and $E_{\rm agg} = 0.5$ eV, as shown in Fig. 4. The same radius is used for both p-Cr_5Si_3 and p-$CrSi_2$ for simplicity.

The resistivity of the thin film which comprises a-CrSi, p-Cr_5Si_3, p-$CrSi_2$, and Si is calculated using the effective medium approximation for conduction in two dimensions[6]. In order to include the effect of grain boundary, the resistivity of a polycrystalline material is expressed as a sum of a crystalline part and a grain boundary part: $\rho_{\rm p} = \rho_{\rm c} + \rho_{\rm gb}$. The grain boundary part is assumed to be inversely proportional to the grain radius, which is a simplified Mayadas-Shatzkes model[7]. The same grain boundary resistivity is used for both p-Cr_5Si_3 and p-$CrSi_2$ for simplicity.

III. Results and Discussion

The activation energies of the transitions and the resistivities and the TCRs of each material are optimized to minimize the root mean square error between measured and calculated resistivities and TCRs.

The extracted values for $E_{\rm a1}$, $E_{\rm a2}$, and $E_{\rm a3}$ are $2.44\,{\rm eV}$, $6.36\,{\rm eV}$, and $2.52\,{\rm eV}$, respectively. The rate of transitions with the extracted activation energies as shown in Fig. 2 suggests that meta-stable crystals appear first, which is known as Ostwald step rule[3]. The meta-stable state remains within a certain range of annealing temperature as shown in Fig. 3. Note that $E_{\rm a1}$ and $E_{\rm a3}$ are close to the measured activation energy $2.6\,{\rm eV}$ for silicide formation in multi-layered thin films of chromium and amorphous silicon[8].

The calculated resistivities and the TCRs with the values of the parameters in Table I are in good agreement with measured data as shown in Figs. 5 and 6. The phase transitions from as-deposited amorphous to poly-crystalline including a meta-stable state lead to high resistivity within a certain range of annealing. The gradual increase of TCR above $600^\circ{\rm C}$ comes from the grain growth which is shown in Fig. 4. The relation between the resistivity and the TCR with various annealing temperatures is also well expressed by the proposed model as shown in Fig. 7.

IV. Conclusion

The electrical properties of CrSi thin films were modeled considering phase transitions and grain growth during an-

nealing. The phase transitions from as-deposited amorphous to poly-crystalline including a meta-stable state lead to high resistivity within a certain range of annealing temperature.

References

[1] N. Ito *et al.*, *ADMETA Plus*, 5-6, 2021.
[2] C. Gladun *et al.*, *Int. J. Electronics*, p. 301, 1994.
[3] R. A. Van Santen, *J. Phys. Chem.*, p. 5768, 1984.
[4] A. Soleimani-Dorcheh and M. C. Galetz, *Oxid. Met.*, p. 73, 2015.
[5] A. Redaelli *et al.*, *J. Appl. Phys.*, 111101, 2008.
[6] S. Privitera *et al.*, *IEEE Trans. Electron Devices*, p. 3549, 2012.
[7] A. F. Mayadas *et al.*, *Appl. Phys. Lett.*, p. 345, 1969.
[8] T. E. Schlesinger *et al.*, *Appl. Phys. Lett.*, p. 449, 1991.

TABLE I
THE RESISTIVITY AND TCR OF EACH MATERIAL NORMALIZED BY THE ABSOLUTE VALUES OF THE AS-DEPOSITED SAMPLE. THE GRAIN RADIUS r_g IS IN NANOMETERS.

material	normalized ρ	normalized TCR
a-CrSi	1.09	-0.768
Cr_5Si_3	0.459	12.4
$CrSi_2$	0.725	3.04
Si	19.4	-3.41
grain boundary	$0.258/r_g$	-13.1

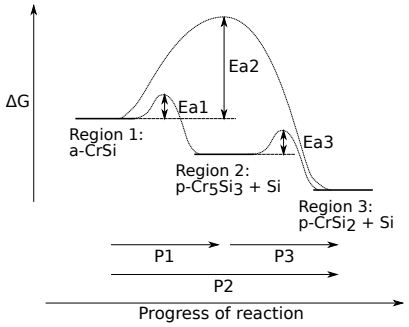


Fig. 1. The schematic diagram of the phase transitions during annealing.

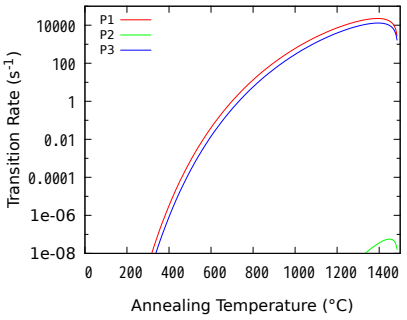


Fig. 2. The temperature dependence of the rate of the transitions shown in Fig. 1.

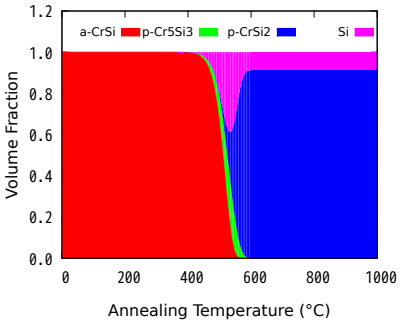


Fig. 3. The volume fraction of each phase. The annealing time is 10 min.

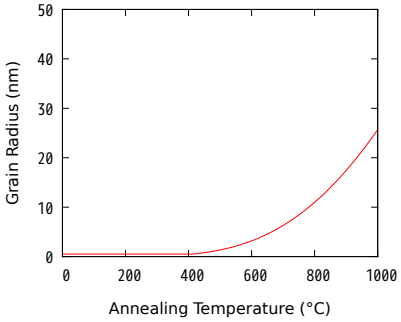


Fig. 4. The grain radius as a function of the annealing temperature. The annealing time is 10 min.

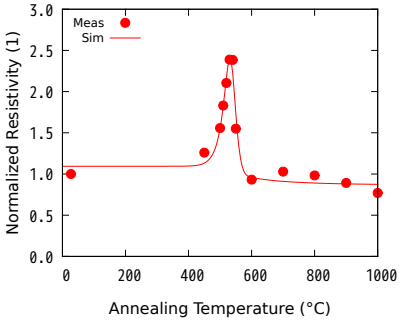


Fig. 5. The resistivity of the thin film as a function the annealing temperature. The resistivity is normalized by the value of the as-deposited sample.

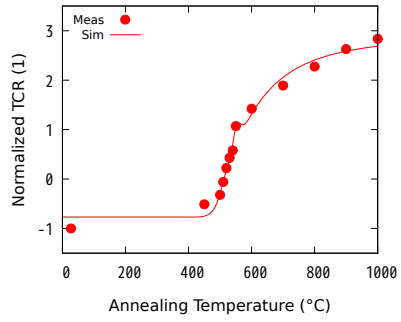


Fig. 6. The TCR of the thin film as a function the annealing temperature. The TCR is normalized by the value of the as-deposited sample.

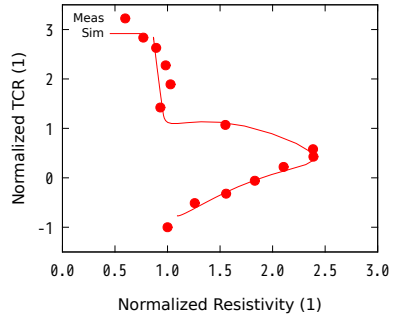


Fig. 7. The dependence of resistivity and TCR on annealing temperature.

Modeling Non-Ideal Conformality during Atomic Layer Deposition in High Aspect Ratio Structures

Luiz Felipe Aguinsky*, Frâncio Rodrigues*, Xaver Klemenschits†, Lado Filipovic†,
Andreas Hössinger‡, and Josef Weinbub*
*Christian Doppler Laboratory for High Performance TCAD at the
†Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, 1040 Wien, Austria
‡Silvaco Europe Ltd., Compass Point, St Ives, Cambridge, PE27 5JL, United Kingdom
Email: aguinsky@iue.tuwien.ac.at

***Abstract*—Atomic layer deposition (ALD) is a key technology in semiconductor processing, enabling precise control over the deposited film thickness and providing high conformality. A fundamental ALD process is the deposition of aluminum oxide (Al_2O_3) from sequential trimethylaluminum (TMA) and water (H_2O) exposure during sequential ALD cycles, having found application in the deposition of high-κ capacitor films for dynamic random-access memory (DRAM). DRAM requires ALD on high aspect ratio structures, where non-ideal conformality can occur. We present an extension to existing one-dimensional flux and surface coverage models by including reversible adsorption. Our approach allows to focus on the H_2O step, which is known to yield non-ideal conformality. We show good agreement of our modeling approach with experimentally reported H_2O-limited thickness profiles. By incorporating these advanced models in a level-set based topography simulator, ALD can be combined with other processing steps to model a complete fabrication sequence. Moreover, the clean surface sticking probabilities derived with our model are consistent with measured values.**

Introduction Atomic layer deposition (ALD) is a thin film growth technique and is an essential enabler of high aspect ratio (HAR) structures such as dynamic random-access memory (DRAM) [1]. The success of ALD comes from its good control over film conformality and thickness. This is achieved by dividing the growth process into self-limiting processing steps, which then repeat in cycles [2]. Thus, film thickness is controlled via the growth per cycle (GPC) and the total number of cycles (N_{cycles}). Ideally, the self-limiting nature of the reactions provides perfect conformality. However, in HAR structures, deviations from ideal conformality are frequently observed due to reactant transport being more restricted as it traverses towards the bottom of the structure.

The focus of our work is ALD of Al_2O_3 from sequential cycles of exposure to trimethylaluminum (TMA) and water (H_2O) [3], which has found applications as high-κ capacitor films for DRAM [1]. The first step is a reaction between ambient TMA and a hydroxyl-terminated (OH-terminated) surface. After purging, ambient H_2O is introduced during the second step, which reacts with the adsorbed TMA, leaving an OH-terminated surface once again. The reaction pathways of the H_2O step are depicted in Fig. 1: Adsorption-reflection, represented by a coverage-dependent sticking probability $\beta(\theta)$, and desorption, modeled using an evaporation flux Γ_{ev}.

Such processes are commonly modeled via first-order Langmuir adsorption and have been applied to predict saturation times [4], to model growth kinetics [5], and to estimate the clean surface sticking probability β_0 [6]. However, they do not directly evaluate the thickness profiles. Such profiles are necessary if, e.g., ALD modeling is to be combined with other semiconductor processing steps and if a process-aware simulation of the device operation is desired. Previously, we tackled this issue by developing a model for ALD of titanium compounds, including complex surface models and a precise evaluation of local fluxes using Monte Carlo ray tracing [7]. However, this model incurs high computational costs.

In this work, we propose a one-dimensional (1D) model for ALD in HAR structures and apply it to the growth of Al_2O_3 in the H_2O-limited regime. The proposed model accurately captures the non-ideal conformality by including reversible Langmuir kinetics [8]. The model is integrated in the level-set based [9] topography simulator ViennaLS [10] in order to calculate the evolution of the thickness profiles.

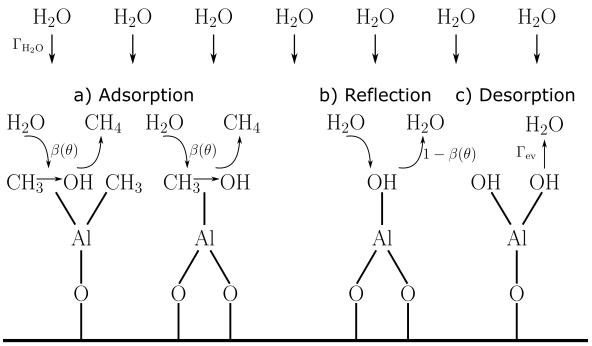


Fig. 1: Possible reaction pathways during the H_2O step.

Method In our 1D ALD growth model, illustrated in Fig. 2, we focus on two main variables: The surface coverage θ and the local reactant flux Γ for a single limiting species [4]. We assume that the complementary species (TMA) has a sufficiently high dose to fully cover the surface. In the reference experiment [6], the applied TMA dose is 1.6 times that of H_2O, thus the limiting factor is the H_2O flux $\Gamma_{\mathrm{H_2O}}$.

Our surface model is an extension of the conventional single-particle Langmuir model [4] with reversible kinetics (i.e., including the evaporation flux Γ_{ev}) to describe the complex H_2O step with improved accuracy [8]. The H_2O flux is calculated using a 1D Knudsen diffusion model for a trench with opening d and depth L [5]. That is, we assume

a preferential transport direction z along the trench sidewall. The equations for θ and Γ_{H_2O}, assuming a surface site area s_0, are provided in (1) and (2) respectively:

$$\frac{d\theta}{dt} = s_0 \overbrace{\beta_0(1-\theta)}^{\beta(\theta)} \Gamma_{H_2O} - s_0 \theta \Gamma_{ev}, \quad (1)$$

$$\frac{d^2\Gamma_{H_2O}}{d^2 z} = \frac{3}{(2\cdot d)^2} \beta_0 (1-\theta) \Gamma_{H_2O}. \quad (2)$$

Eq. (1) is solved with Euler's method with a pulse time t_p and Eq. (2) is solved with a central finite differences scheme.

A crucial challenge is that ALD does not operate in steady-state, but in cycles. We tackle this by introducing an artificial time $t^* = N_{cycles}/C$ during which the film is grown, where C is a numeric constant. The film grows in a direction normal to the surface element at position $\vec{r}$ with a growth rate $GR(\vec{r})$ given by:

$$GR(\vec{r}) = GR(z) = C \cdot \mathrm{GPC} \cdot \theta(z). \quad (3)$$

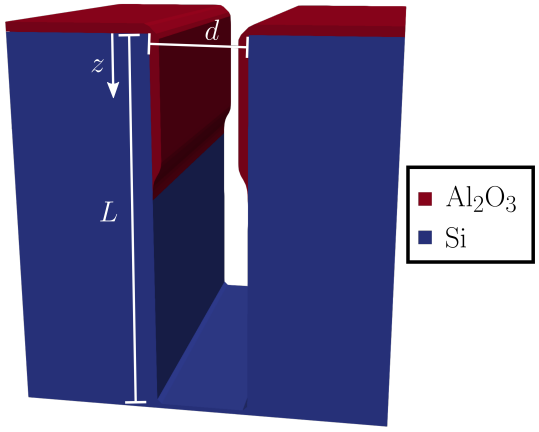


Fig. 2: Illustration of the ALD model simulation domain.

Results We calibrate our model to measured profiles of Al_2O_3 reported by Arts *et al.* [6]. The measurements were made in a lateral HAR structure ($d = 0.5\,\mu$m, $L = 5\,$mm) tailored to film conformality analyses [11]. A maximum thickness of 46 nm was achieved in the H_2O-limited regime ($\approx$750 mTorr·s) after 400 ALD cycles for a GPC of 1.15 Å at three substrate temperatures (150 °C, 220 °C, and 310 °C).

We model the lateral trench in two dimensions (yz), where y is the direction of film growth and z is the lateral coordinate and transport direction. From (1) and (2), we calculate the model parameters Γ_{ev} and β_0, whose calibrated values are shown in Table I. The profiles are reported in Fig. 3.

TABLE I: Parameters calibrated to measurements from [6].

Parameter	150 °C	220 °C	310 °C
Γ_{ev} ($m^{-2}s^{-1}$)	$6.5\cdot10^{19}$	$5.5\cdot10^{19}$	$3.5\cdot10^{19}$
β_0	$5.0\cdot10^{-5}$	$1.6\cdot10^{-4}$	$1.9\cdot10^{-4}$
β_0, estimated range from [6]	$1.4\cdot10^{-5}$ – $2.3\cdot10^{-5}$	$0.8\cdot10^{-4}$ – $2.0\cdot10^{-4}$	$0.9\cdot10^{-4}$ – $2.5\cdot10^{-4}$

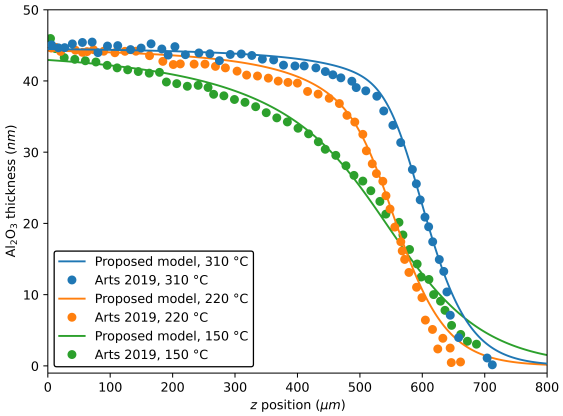


Fig. 3: Comparison of calibrated simulation to thickness profiles measured by Arts *et al.* [6].

Discussion In Table I we observe that the calibrated values of β_0 are generally consistent with those estimated in [6]. This is expected, since the methodology used for those estimations is also based on first-order Langmuir kinetics. Nonetheless, we expect that our approach provides a more accurate estimate, particularly for the discrepant value at 150 °C, since we consider the entire thickness profile instead of the slope at 50% height and we include an evaporation flux Γ_{ev}.

From Fig. 3, we note that our modeling approach provides a good agreement with the reported profiles. This agreement, combined with the measurable variation of maximum thicknesses with temperature, is a strong indicator that the temperature dependence of the evaporation flux Γ_{ev} plays a key role in determining the profile. Therefore, Γ_{ev} must be taken into account for the accurate modeling of ALD.

Acknowledgments

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

This work was supported in part by the Austrian Research Promotion Agency FFG under Project 878662 PASTE-DTCO.

References

[1] K. Seshan *et al.*, *Handbook of Thin Film Deposition*. Elsevier, 2018.
[2] V. Cremers *et al.*, *Applied Physics Reviews*, vol. 6, no. 2, p. 021302, 2019.
[3] S. M. George, *Chemical Reviews*, vol. 110, no. 1, pp. 111–131, 2010.
[4] A. Yanguas-Gil, *Growth and Transport in Nanostructured Materials: Reactive Transport in PVD, CVD, and ALD*. Springer, 2016.
[5] M. Ylilammi *et al.*, *Journal of Applied Physics*, vol. 123, no. 20, p. 205301, 2018.
[6] K. Arts *et al.*, *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 37, no. 3, p. 030908, 2019.
[7] L. Filipovic, in *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 323–326, 2019.
[8] B. A. Sperling *et al.*, *The Journal of Physical Chemistry C*, vol. 124, no. 5, pp. 3410–3420, 2020.
[9] J. A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press, 1999.
[10] "ViennaLS." [Online]. Available: https://viennatools.github.io/ViennaLS
[11] F. Gao *et al.*, *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 33, no. 1, p. 010601, 2015.

A Novel Methodology for Neural Compact Modeling Based on Knowledge Transfer

Ye Sle Cha, Junghwan Park, Chanwoo Park, Soogine Chong, Chul-Heung Kim, Chang-Sub Lee, Intae Jeong and Hyunbo Cho
Research & Development Center, Alsemy Inc., Seoul, Korea. Email: yesle.cha@alsemy.com

***Abstract*—This work presents a novel approach of using knowledge transfer to increase the accuracy of artificial neural network (ANN)-based device compact models, or neural compact models. This is useful when the amount of data available for training an ANN is limited. By utilizing relatively abundant data of a previous technology node, physical phenomena that are not evident in the limited data of the target technology node (e.g. gate-induced drain leakage) are accurately predicted. When meta learning algorithms are used, the accuracy of the model significantly increases, with relative linear error 10 times lower compared to the case when prior knowledge is not incorporated. The proposed methodology can be used to model future generation devices with limited data, utilizing data from well-characterized past technology node devices.**

***Index Terms*—Artificial neural network, compact modeling, deep learning, knowledge transfer, meta learning, MOSFET, statistical modeling, transfer learning.**

I. Introduction

Artificial neural network (ANN)-based device compact models, or neural compact models, use machine learning to model devices for circuit simulation based on data [2]. However, it is often difficult to obtain a large dataset from novel devices, which limits the accuracy of the neural compact model. To overcome this issue, previous work directly incorporated in the ANN model the physics that was already understood [1]. In this paper, we propose knowledge transfer methods in which the relevant physics in the data is automatically included in the model, without the need to understand such physics.

II. Knowledge Transfer For Device Modeling

The objective of this paper is to provide a new modeling framework for tackling the scarcity of data for a target device. In this framework, we set a similar environment to the case when fitting parameters are extracted for analytical models (e.g. BSIM), where only few I-V sweeps are measured for each channel width (W), channel length (L), and temperature (T) of the target device. Fig. 1 shows the contrast between available data, which are used for ANN training, and test data, which are used to examine the accuracy of the final ANN.

The methodology consists of two parts. First, we *pretrain* the ANN to learn device physics from largely available planar MOSFET data. For each W, L, T, a large amount of data, on par with the union of two data shown in Fig. 1, is used to capture the delicate physical phenomena. Second, we *adapt* the pretrained ANN to the limited target device data, so that the learned physics becomes consistent with the target device. The whole procedure is called 'knowledge transfer'. The details of how each part is conducted depend on the knowledge transfer method used, as explained in Fig. 2.

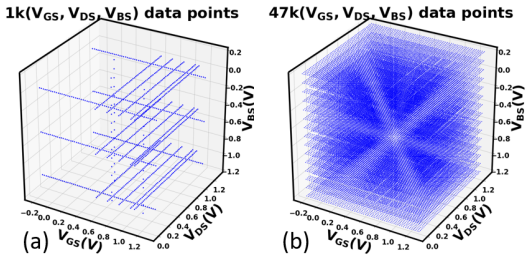


Fig. 1. Bias domain for each W, L, T of the target device of (a) available data, consisting of 24 current-voltage sweeps, and (b) test data with approximately 47 times more data points than those of the available data.

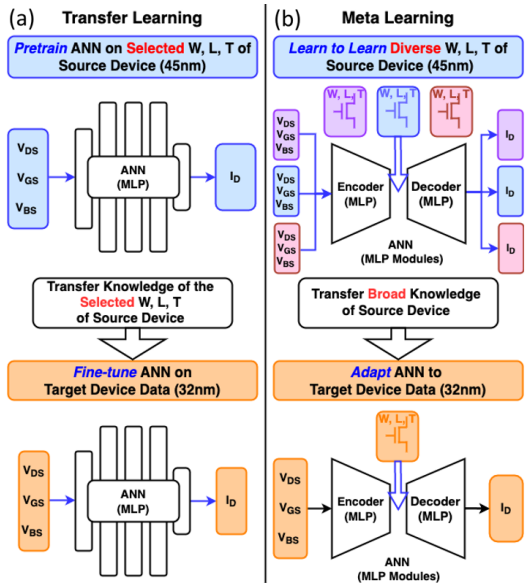


Fig. 2. (a) For transfer learning, we *pretrain* a multilayer perceptron (MLP) on a selected W, L, T dataset of the source device, then *fine-tune* that ANN on limited W, L, T data of the target device. (b) For meta learning, the ANN *learns to learn* by being *meta trained* to repeatedly make accurate predictions utilizing the condensed information on each W, L, T data of the source device that the encoder extracts (represented by transistor symbols). The ANN is adapted to the target device data by extracting relevant information with limited W, L, T data (transistor symbol), and instantiates predictions (in black).

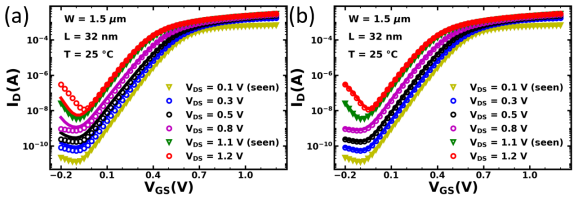


Fig. 3. I_D-V_{GS} ANN predictions (lines), training data (triangles), and test data (circles) of (a) the randomly initialized ANN and (b) the fine-tuned ANN, showing the fine-tuned ANN accurately modeling GIDL.

TABLE I
KNOWLEDGE TRANSFERED ANN VS. RANDOMLY INITIALIZED ANN

	Random Initialization	Transfer Learning	Meta Learning
Pretraining Time	N/A	646 sec.	17 hours
Adaptation Time (per W, L, T)	538 sec.	186 sec.	1 sec.
Relative Linear Error (%)	22.9	4.3	2.3
Relative Log Error (%)	1.56	0.40	0.11

While transfer learning transfers knowledge of a selected W, L, T dataset of the source device, with meta learning, we aim to build a machinery which effectively passes broad knowledge of *all* W, L, T datasets of the source device. In meta learning, the ANN *learns to learn* by performing several learning episodes with diverse W, L, T datasets as described in Fig. 2. After meta training is completed, the ANN quickly leverages the information from limited data of *any* W, L, T of the target device to instantiate accurate predictions on large test data. We also apply MetaFun techniques [3] to capture subtle interactions between data points.

III. EXPERIMENTS

To test the proposed framework, 45 nm and 32 nm technology node MOSFETs [4] are used as source and target devices, respectively. Data for experiments are generated by SPICE simulations. We validate our methodology by comparing it to the case where the ANN is randomly initialized and trained solely on the limited data of the target device.

In Fig. 3, in contrast to the randomly initialized ANN, the fine-tuned ANN successfully predicts gate-induced drain leakage (GIDL) current by using previously learned physical knowledge. Similarly, in Fig. 4, the meta trained ANN predicts I-V characteristics more accurately for up to two differentiations compared to the randomly initialized ANN, without any pre-imposed conditions on derivatives.

Table I compares computational costs and test errors for all three ANN training methods. During pretraining, one W, L, T dataset is used for transfer learning, and 240 such datasets are used for meta learning, with a long training time. For adaptation to the target device data, 52 W, L, T datasets are used for all methods. The meta trained ANN shows the lowest average test errors with the shortest adaptation time.

In Fig. 5, the test results for matching electric parameters such as I_{DLIN}, I_{DSAT}, V_{TH}, and GIDL current are shown for each ANN training method. The meta trained ANN captures the electric parameters for *any* W, L, T in a much more stable and accurate way compared to the randomly initialized ANN. The fine-tuned ANN shows good performance, but its limit lies on higher variance.

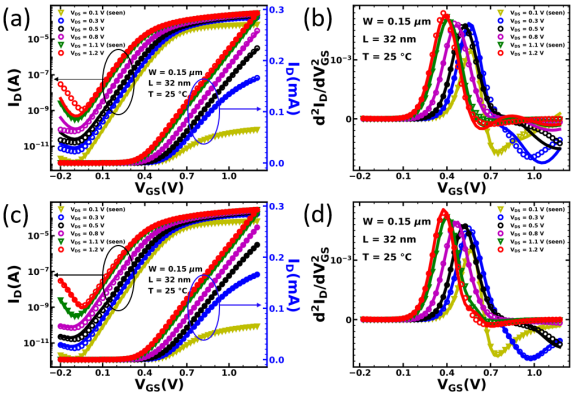


Fig. 4. I_D-V_{GS} and second derivative of I_D-V_{GS} ANN predictions (lines), training data (triangles), and test data (circles) of the randomly initialized ANN [(a) and (b), respectively], and of the meta trained ANN [(c) and (d), respectively], showing the meta trained ANN achieving significantly better fitting compared to the randomly initialized ANN.

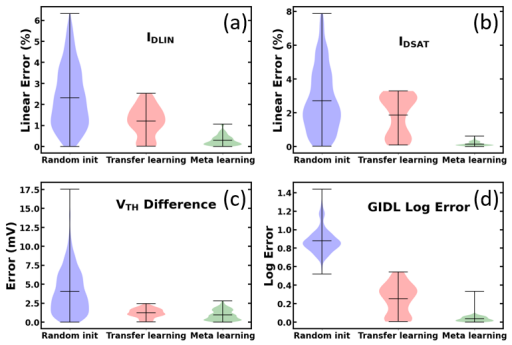


Fig. 5. Comparison of relative linear errors for fitting (a) I_{DLIN} and (b) I_{DSAT}, (c) V_{TH} differences, and (d) log errors for fitting GIDL current by ANN predictions for three methods, showing improved model accuracy for knowledge transfer methods compared to random initialization. For target device data, we select 108 well-balanced W, L, T datasets and reduce the available data for each W, L, T by nearly 36 percent compared to Fig. 1 (a).

IV. CONCLUSION

We develop a novel framework for neural compact modeling by applying advanced knowledge transfer techniques. The resulting model learns the device physics underlying widely available planar MOSFET data, and uses that knowledge to predict physically consistent I-V characteristics for any W, L, T of a target device with excellent accuracy, even if the available data of that device are limited.

REFERENCES

[1] Y. Kim et al., *SISPAD*, pp. 257-260, 2020.
[2] J. Wang et al., *IEEE Trans Electron Devices*, vol. 68, no. 3, pp. 1318-1325, 2021.
[3] J. Xu et al., In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10617-10627, 2020.
[4] Predictive Technology Model (PTM), http://ptm.asu.edu/

Graph-based Compact Modeling (GCM) of CMOS Transistors for Efficient Parameter Extraction: A Machine Learning Approach

Amol D. Gaidhane[1], Ziyao Yang[1], and Yu Cao[1]
[1]*School of ECEE, Arizona State University, Tempe, AZ, USA*
Email: yu.cao@asu.edu / Phone: 480-965-1472

Abstract
Parameter extraction of compact transistor models is an expensive process, heavily relying on engineering knowledge and experience. To automate such a process, we propose a novel approach, Graph-based Compact Model (GCM), that integrates physical modeling and data-driven learning. GCM utilizes Graph Neural Networks (GNNs) to establish the model structure, while retaining the physicality in compact models. We implement our GCM in Verilog-A to support circuit simulations. As demonstrated with an academic 7nm FinFET PDK [1], the new approach automatically generates a GCM model within a minute, and achieves excellent accuracy and efficiency in SPICE.

Introduction
To capture the complexity of device physics, an increasing number of model parameters have to be introduced, posing enormous challenges in parameter extraction and simulation efficiency. Artificial neural networks were able to fit transistor behaviors [2]. Without explicit physical meaning, such ANNs impede model scalability and efficiency.

Recently GNNs were proposed to model dynamic systems [3]. GNNs describe each physical component as the graph node, and their interactions as edges. Based on GNNs, we introduce our new method, GCM, for compact transistor modeling. For key physical parameters, they are captured by non-linear models and embedded into the graph nodes. For many other fitting parameters, they are replaced by neural networks to connect the nodes together. Parameter extraction, i.e., training of GCM, is data driven through back propagation, with appropriate constraints on physical parameters to improve the robustness.

GCM Model Development
The new method of GCM converts a conventional model into a set of nodes for key physical parameters, and connects them into a directional graph. Fig. 1(a) presents the graph structure to model a transistor, with nodes for selected physical parameters and edges for their relationships. Similar as other compact transistor models, GCM receives the input features as an vector (e.g., voltages, geometries, etc.), and predicts the output features (e.g., $I_{\rm ds}$ and its derivatives).

There are two basic operations in GCM, aggregation and transformation. The aggregation step computes the node value from the input vector or its nearest neighbors, depending the graph structure. The transformation step applies the update function to generate a new value for each node and the output vector. The update function can be physical equations, if the physics is clear, or neural networks for fitting. We keep as many nonlinear relations as possible in GCM to minimize the model size.

To demonstrate GCM, we start from a long-channel surface-potential based model of FinFET [4]. β_S and β_D are the intermediate nodes which are solved after applying boundary conditions to the Poisson's equation. $\Delta\phi$ is a node for the work-function difference. Moreover, for the short channel effects (SCEs), we add certain nodes, such as n for sub-threshold slope. Similarly, $\Delta V_{\rm th}$ is a node for the change in threshold voltage, Moc is for channel length modulation (CLM), Mu is for the effective mobility, and $V_{\rm dsat}$ is for the drain saturation voltage, etc. Overall, the nodes preserve important physics for the FinFET transistor.

Fig. 1(b) presents an example of GCM feature transformation for the node n, which is a combination of physical equations and a multi-layer perceptron (MLP) to fit the $V_{\rm ds}$ dependency. As $n_{\rm L_{eff}}$ has a specific dependence on the channel length ($L_{\rm eff}$) [5], we adopt the compact model to keep the physicality and minimize model fitting.

Fig. 2 shows a two-phase training procedure for efficient extraction of the GCM model, balancing multiple objectives in the loss function. In Phase 1, we use $I_{\rm ds}$-$V_{\rm gs}$ data to train the model, with G_m in the loss function to improve the accuracy. GCM extracts $V_{\rm ds}$ independent parameters in this phase. In Phase 2, we freeze the learned parameters in Phase 1 and extract $V_{\rm ds}$ dependent parameters. We use IV data generated from a BSIM-CMG based FinFET model at 7nm [1] to train the GCM model. Fig. 3 validates main electrical characteristics of the n-FinFET and p-FinFET between the BSIM-CMG model and GCM. The GCM model accurately captures the drain current and it's derivatives for both n-FinFET and p-FinFET.

Fig. 4 illustrates the training curves for several GCM model parameters. For the 7nm FinFET, GCM has 113 parameters and extracts all model parameters in 31.33 seconds. This is significantly faster than the conventional extraction process, which usually takes hours to days. Fig. 5 demonstrates circuit simulation with GCM through Verilog-A. GCM achieves similar accuracy as BSIM, with shorter simulation time in SPICE.

Conclusion
We propose a novel graph-based compact model which is physical and efficient in parameter extraction. The graph structure preserves key physical models, supports automatic learning from data, and is flexible to incorporate more advanced effects (e.g., cryogenic effects).

References
[1] L. T. Clark et al., J. Microelectronics, 53, 2016.
[2] L. Zhang et al., J Comput Electron, 16, 2017.
[3] S. Kazemi et al., J. Mach. Learn. Res., 21, 2020.
[4] Y. Taur et al., IEEE EDL, 25, 2004.
[5] Y. S. Chauhan et al., Academic Press, 2015.

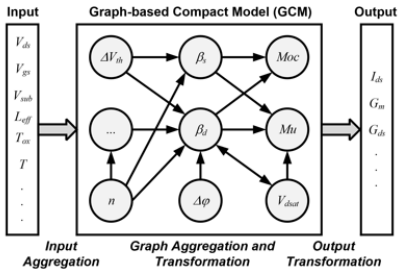


Fig. 1(a): The structure of graph-based compact model.

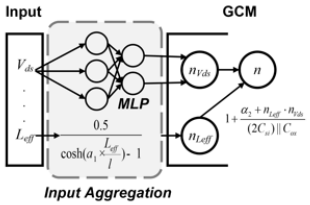


Fig. 1(b): Combination of MLP and models in aggregation.

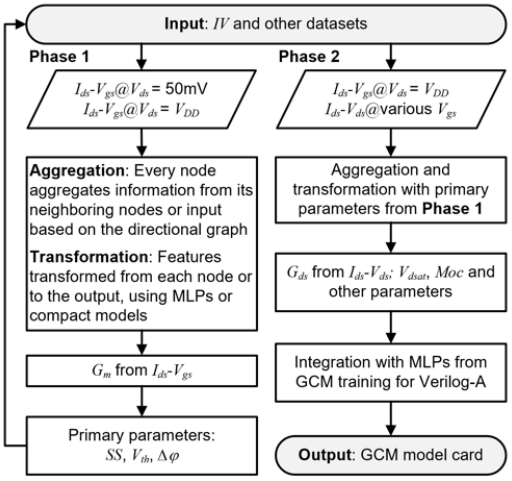


Fig. 2: A two-phase training procedure for GCM.

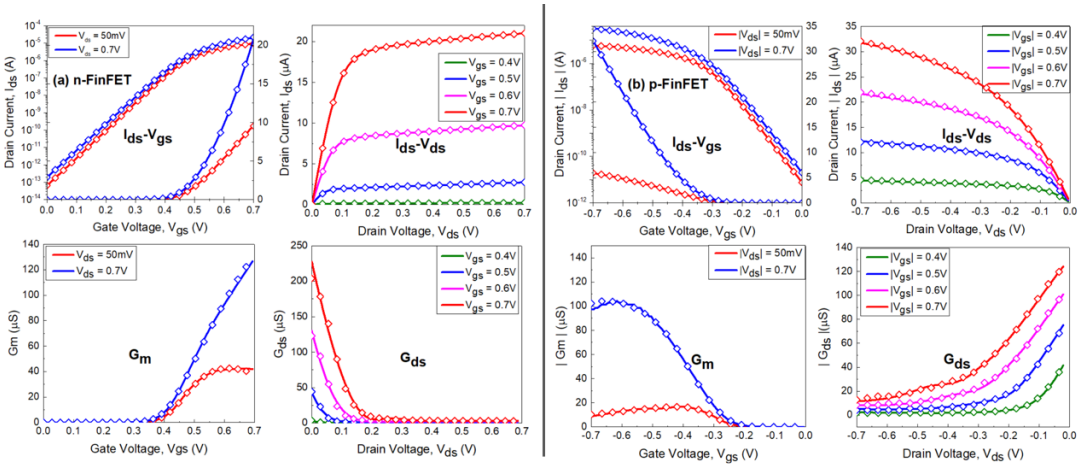


Fig. 3: Validation of the GCM model (line) with a 7nm BSIM-CMG model (symbols), in the prediction of I_{ds}, G_m and G_{ds} for (a) n-FinFET and (b) p-FinFET.

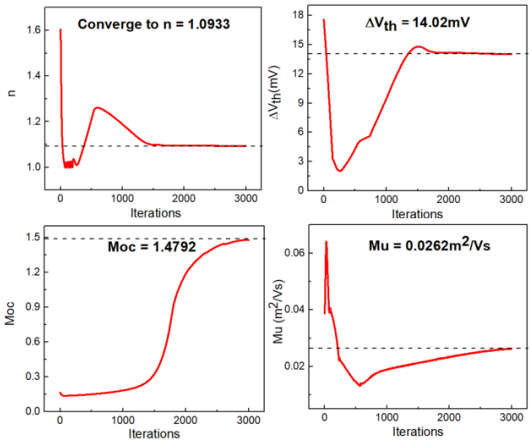
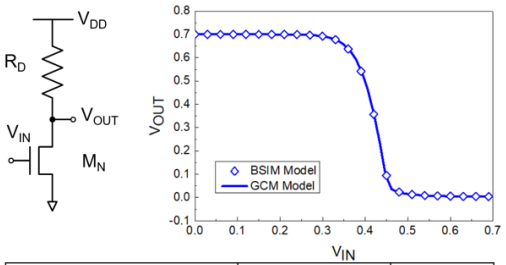


Fig. 4: The training curve of several GCM parameters. It only takes **31.33 seconds** to complete the training of GCM.



	BSIM	GCM
SPICE Simulation Time (ms)	12.745	5.284
Number of Model Parameters	905*	113

*This is for a complete BSIM-CMG model card

Fig. 5: Comparison of circuit simulation of GCM with BSIM-CMG in SPICE.

Hierarchical Mixture-of-Experts Approach for Neural Compact Modeling of MOSFETs

Chanwoo Park, Premkumar Vincent, Soogine Chong, Junghwan Park, Ye Sle Cha, and Hyunbo Cho
Research & Development Center, Alsemy Inc., Seoul, Korea. Email: chanwoo.park@alsemy.com

***Abstract*—With scaling, physics-based analytical MOSFET compact models are getting more complex, and the number of fitting parameters has increased immensely. Parameter extraction based on measured or simulated data consumes a significant time in the compact model generation process. We present a Mixture-of-Experts approach to neural compact modeling that can reduce the compact model generation time while being robust enough to be applicable to upcoming future CMOS technology. Our results show that the proposed method is 30.6% more parameter-efficient and achieves higher accuracy using fewer data when compared to conventional neural compact modeling.**

***Index Terms*—neural compact model, mixture-of-experts, artificial neural network, MOSFET**

I. Introduction

Compact modeling acts as a bridge between device fabrication and circuit design. It has two main goals: computational efficiency and accuracy. Conventionally, to achieve such contradicting goals, analytical approximations and empirical fitting parameters have been used. Threshold voltage-based models had issues in harmonic distortion analysis which took a considerable time for surface-potential-based models to solve [2]. Since the conventional compact models are technology dependent, it could take years to develop new models for upcoming devices. Hence, there is an immediate need for fast and efficient compact model generation. Our approach to dwell past this obstacle is artificial neural network (ANN)-based compact modeling, or neural compact modeling; however, Simple "black-box" neural compact models are computationally inefficient, non-robust, and require a large dataset to work on.

Our proposed Mixture-of-Experts (MoE) offsets the above limitations by sub-categorizing the problem into different operation regimes and solving each regime using a dedicated expert ANN. In this article, we apply the MoE approach to neural compact modeling based on the fact that the MOSFET characteristics can be segregated into sub-regions which share similar physical properties.We also analyze the parameter-efficiency and the accuracy of our MoE model over the conventional neural compact model.

II. Mixture-of-Experts for Neural Compact Modeling

The Mixture-of-Experts is based on the *divide and conquer* principle, which was first introduced in [1]. The MoE works on the idea that the whole input space can be partitioned into smaller regimes of distinct characteristics. For accurate MOSFET modeling, all its sub-characteristics such as gate-induced drain leakage, short-channel effect, etc. need to be modeled. Using a single large ANN to model them is challenging. Thus, we address this problem by partitioning the input domain and assigning specialized local models for each partition.

The entire network consists of three-level mixture of experts as shown in Figure 1. At each level, the outputs of the expert networks are provided with importance weights by their gating network. Note that all these processes are trained *automatically* in an end-to-end manner with a mean-squared-error loss function. The gating network learns to assign larger weight to the more appropriate expert based on the input vector to improve the performance of the entire network and the experts gradually learn to focus more on their own specialized characteristic regions for efficient training. In our experiments, all experts and gating networks are simple feed-forward neural networks, where the output of gating networks is softmax assigning mixing weights to each expert.

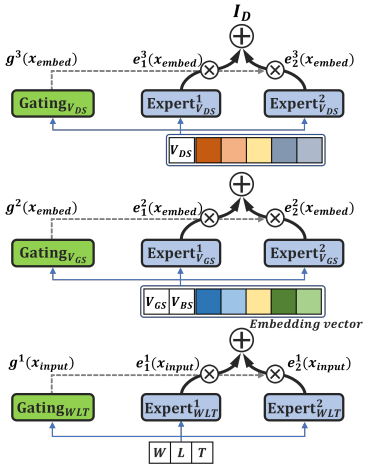


Fig. 1: Mixture-of-Experts based neural compact modeling. The gating network categorizes the input vector and assigns weights to the output of each expert network. The weighted sum of the outputs is passed on to the next level.

We denote $g^k(x)_i$ and $e_i^k(x)$ as the output of the gating network and the output of the i-th expert network at the k-th level for a given input vector x, respectively. The final embedding vector $f^k(x)$ at each level is the linearly weighted sum of each expert's output, expressed as follows:

$$f^k(x) = \sum_{i=1}^{N} g^k(x)_i e_i^k(x) \tag{1}$$

where N is the number of experts at the same level and $\sum_i g^k(x)_i = 1$.

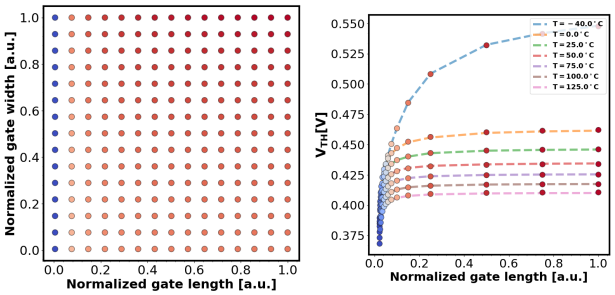
The final embedding vector is passed on to the next gating network where different experts take control for a different input regime.

III. Experiments and Results

Our dataset was generated from the SPICE simulation of 45nm PTM HP modelcard [3]. In our discussion, $e_1^k(x)$ is denoted by blue, and $e_2^k(x)$ by red. Intermediate shades represent a mix of both experts.

A. WLT Gating

At the WLT gating network, gate width (W), gate length (L), and temperature (T) characteristics were sub-categorized



(a) (b)

Fig. 2: Representation of categorized regions: (a) W vs L and (b) V_{TH} vs L for short and long gate length regimes

by the experts and modeled. $e_1^1(x)$ expertised in short-channel effects while $e_2^1(x)$ took over long channels as seen from Figure 2a. Figure 2b shows threshold voltage (V_{TH}) vs varying gate length where the decreased short-channel V_{TH} was modeled primarily by $e_1^1(x)$ while the long-channel V_{TH} was modeled by $e_2^1(x)$. The reverse narrow-width effect was negligible and, thus, a separate expert was not required.

B. V_{GS} Gating

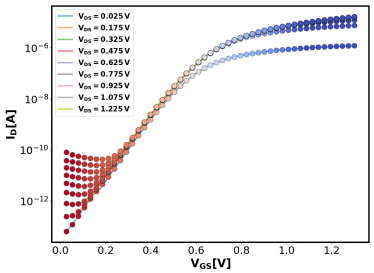


Fig. 3: Categorized I_D-V_{GS} regions: $I_D \propto \exp(V_{GS})$ in OFF region, and $I_D \propto V_{GS}$ in ON region

The embedding vector from the WLT gate was combined with V_{GS} and V_{BS} to create the input vector for the V_{GS} gating network. $e_1^2(x)$ was designed to take V_{GS} as an input to model the approximately linear or quadratic V_{GS} dependence of I_D in the ON-state, while $e_2^2(x)$was designed to take $\exp(V_{GS})$ as the input to model the exponential V_{GS} dependence of I_D in the sub-threshold region (figure 3). The transition region exhibited a smooth continuous change in prioritizing the expert based on the operation region.

C. V_{DS} Gating

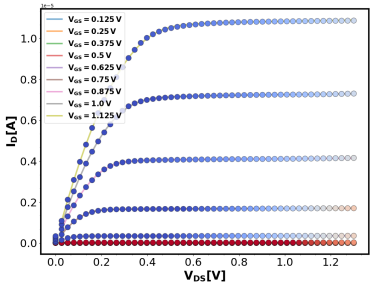
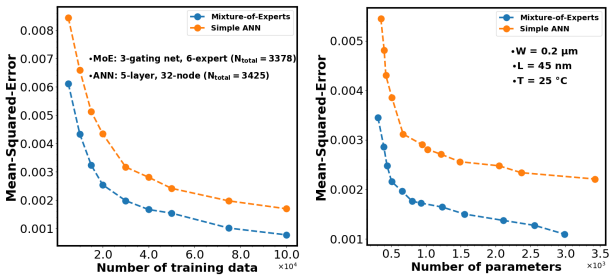


Fig. 4: Categorized I_D-V_{DS} regions: $I_D \propto \exp(V_{DS})$ in the cut-off region, $I_D \propto V_{DS}$ in the linear region, and I_D is weakly dependent on V_{DS} in the saturation region

The final I_D was predicted by the gating network and the experts which were in control of V_{DS} regions, *i.e.*, cut-off, linear, or saturation region. While I_D in the cut-off and the saturation regions do not heavily depend on V_{DS}, the secondary effects introduce a V_{DS} dependency. While the leakage current in the cut-off region shows a $\exp(V_{DS})$ dependency, channel length modulation, drain-induced barrier lowering, and substrate current induced body effect in the saturation region depend on V_{DS}. Hence, the gating network assigned $e_1^3(x)$ to model the cut-off region, $e_2^3(x)$ to model the linear region, while the saturation region was modeled by both the experts (Figure 4).

In our method, the choice of gating decisions was continuous rather than discrete. Any changes of physical phenomena, *i.e.* the linear/saturation regions in the transistor, were not strictly partitioned, but underwent a continuous transition. Therefore, if two experts respectively controlled the linear and the saturation regions, we designed the gating network to be able to continuously mix the two experts rather than discretely select one of them.

IV. Performance Comparison



(a) (b)

Fig. 5: Comparison of test accuracy as a function of (a) the number of training data with the similar model capacity, (b) the number of parameters on the same W, L, T dataset

Figure 5 shows that, with a similar number of parameters (N_{total}), the MoE consistently achieved higher accuracy compared to the simple ANN with a wide range of training data size. Also, the MoE required 30.6% less parameters on average to achieve a similar accuracy as the simple ANN. Fewer parameters in a neural compact model not only result in decreased SPICE computation time, but also achieve higher accuracy when data is limited by better generalization.

V. Conclusion

In this paper, we propose a novel MoE structure for neural compact modeling which utilizes the fact that MOSFETs have distinct characteristics for each input region. A parameter-efficient neural compact model is demonstrated using the MoE structure where we have light-weight experts specialized in each region rather than a single large neural network that has to learn the entire input region. We show that the proposed MoE architecture was more parameter-efficient and achieved higher accuracy using fewer data .

References

[1] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[2] CC McAndrew. Compact modeling: Principles, techniques, and applications. In *Statistical Modeling using Backward Propagation of Variance*. Springer New York, 2010.

[3] Nanoscale Integration and Modeling Group, ASU. PTM High Performance 45nm MOSFET. http://ptm.asu.edu/modelcard/HP/45nm_HP.pm, 2008.

Quantum Element Method for Multi-Dimensional Nanostructures Enabled by a Projection-based Learning Algorithm

Martin Veresko and Ming-C. Cheng*, Dept. of ECE, Clarkson University, Potsdam, NY, USA,
*Corresponding Author: mcheng@clarkson.edu

The solution of the Schrödinger equation in nanostructures is desired in many scientific and engineering applications including materials, medicines, electronics, photonics, [1-8] etc. Such solutions are often computationally intensive, especially when a high accuracy and a fine resolution are needed. This work continues the investigation presented at SISPAD-2021 on an effective quantum simulation methodology for electron wave functions (WFs) in multi-dimensional nanostructures [9]. The methodology was developed based on proper orthogonal decomposition (POD) [10,11] that was applied to project the Schrödinger equation from a nanostructure onto a functional space represented by a finite set of basis functions (or POD modes). The approach is enabled by a learning algorithm via WF data to significantly reduce the degrees of freedom (DoF). Two different projections were studied in the previous work [9], including an individual projection of the WF from each quantum state (QS) and a global projection of the WFs from all selected QSs. The global model appears to be superior to the individual one and offers a very efficient approach to predict the WFs in a 2D quantum dot (QD) structure with a high accuracy [8]. The training to generate the POD modes however requires WF solution data in the nanostructure collected from direct numerical simulations (DNSs) of the Schrödinger equation. For a large-scale multi-dimensional nanostructure, the intensive computing effort to train the POD modes may be prohibitive.

In this work, the POD simulation methodology, together with the quantum element method (QEM), that was studied for 1D quantum wells [12], is extended to multi-element structures to improve effectiveness of the training and simulation of large-scale multi-dimensional nanostructures. Using the QEM, building blocks (or generic *elements*) can be trained and stored in a database for further design/simulation of large nanostructures. Such an approach also empowers effective parallel computing. To enforce the thermal continuity at element interfaces, the discontinuous Galerkin (DG) method [13,14] is applied. The formulation for the QEM is briefly presented below.

The POD QEM generates the modes η from solution data (the WF φ in our case) in each element by maximizing the mean square inner product of η and φ [9,10]. This leads to an eigenvalue problem of a 2-point correlation matrix,

$$\int_{\Omega'} \langle \varphi(\mathbf{r}) \otimes \varphi(\mathbf{r}) \rangle \, \eta(\mathbf{r}') d\Omega' = \lambda \, \eta(\mathbf{r}), \tag{1}$$

where $\otimes$ is a tensor operator, λ is the eigenvalue and the brackets indicate an average over many sets of WF data accounting for variations of potential and boundary conditions (BCs). φ in each element is given as,

$$\varphi(\mathbf{r}) = \sum_{j=1}^{M} a_j(t) \, \eta_j(\mathbf{r}), \tag{2}$$

where M is the selected number of modes for the solution and $a_j(t)$ is the coefficient for the jth mode. To solve a_j, the Schrödinger equation of each element is projected onto a POD space, and a Hamiltonian equation of a multi-element structure for $a_{p,i}$ (the pth element projected along the ith mode) can be written as

$$\sum_{j=1}^{M_p} \left(T_{\eta_p,ij} + U_{\eta_p,ij} \right) a_{p,j} + \sum_{q=1,q\neq p}^{N_{el}} \sum_{j=1}^{M_p} B_{p,pq,ij} a_{p,j} + \sum_{q=1,q\neq p}^{N_{el}} \sum_{j=1}^{M_q} B_{pq,ij} a_{q,j} = E \, a_{p,i} \,, \tag{3}$$

where N_{el} is the total number of elements, M_p and M_q are the selected numbers of modes in the pth and qth elements and the matrix entries of interior kinetic energy $T_{\eta_p,ij}$, potential energy $U_{\eta_p,ij}$, boundary kinetic energies $B_{p,pq,ij}$ and $B_{pq,ij}$ will be given in the extended conference paper.

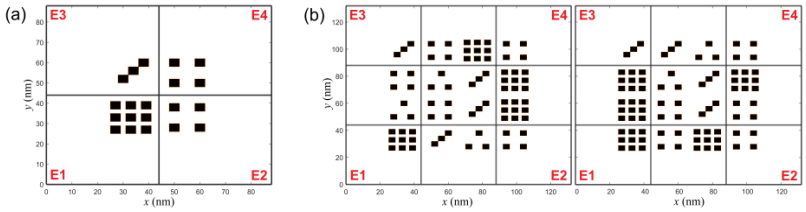


Fig.1. (a) A 4-element structure for demonstration of QEM and (b) two 9-element structures for WF data collection.

The multi-element POD methodology is demonstrated below in a GaAs/InAs QD structure displayed in Fig. 1(a), where the effective mass in GaAs $m^*_{GaAs} = 0.067m_o$ and in InAs $m^*_{InAs} = 0.023m_o$ and the band offset $\Delta E = 0.544$eV. To demonstrate the model construction for the QEM, the structure is first partitioned into 4 subdomains (or elements), E1-E4, shown in Fig. 1(a). DNSs of the Schrödinger equation using a finite difference method for 2 training structures given in Fig. 1(b) are performed to collect WFs in the first 6 QSs for each of 4 elements that are placed in 4 corners of each training structure. 2 groups of orthogonal electric fields in x and y directions are

applied in DNSs, where 6 fields are applied in each direction, varying evenly between –20kV/cm and +20kV/cm. Including the unbiased simulation, the number of WF data samples from 2 structures for the training for each element with 6 QSs in both directions is equal to $(6 \times 2 + 1) \times 6 \times 2 = 156$. The method of snapshots [15-17] is applied to (1) to generate the eigenvalues and POD modes that are then used to evaluate the coefficients in (3).

For demonstration, an electric field of $10\hat{x} - 15\hat{y}$ is applied to the 4-element domain in Fig. 1(a). The least square (LS) errors of the WFs in QSs 1-8 predicted by the POD-QEM, as functions of the number of modes per element, are shown in Fig. 2. The LS error near or below 1% can be achieved in QS 1 with just 2 modes, in QS 2 with 5 modes, in QS 3 with 11 modes, in QS 4 with 10 modes and in QS 6 with 13 modes. In QS 5, the LS error with 10 modes reaches 1.2%. Even for the untrained 7th and 8th states, the LS errors as low as 1.31% and 1.68% are observed when using 13 and 19 modes, respectively, and their errors decline gradually as more modes are added. Contours of $|\varphi|^2$ in QSs 4, 6, 7 and 8 derived from the QEM and DNSs are compared in Fig. 3. The comparison of eigenenergies in QSs 1-8 derived from both approaches are also listed on Table I. These results reveal the accurate prediction of the WFs and eigenenergies by the QEM even for the untrained QSs.

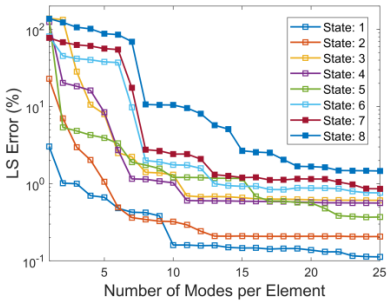


Fig. 2. Least square error in each QS as a function of the number of POD mode.

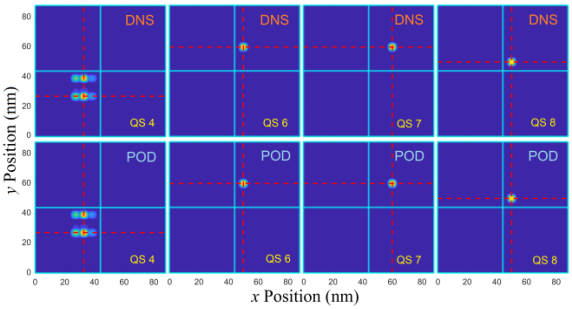


Fig. 3. Contours of $|\varphi|^2$ in QSs 4, 6, 7 and 8. Top row is from the QEM and the bottom on from DNS.

The WFs in QSs 4, 6, 7 and 8 along the paths in x and y directions indicated by the red dashed lines in Fig. 3 are illustrated in Fig. 4, where the WFs from the QEM and DNS are nearly on top of each other once the LS error of the QEM is near or below 2%. For example, in QS 4 an evident deviation in the QEM is observed with 4 modes. Once 7 modes are used, the LS error reduces to 1.16% and WFs from both approaches nearly overlap. Similarly in QS 6, WFs are indistinguishable in Fig. 4 from both approach when a 2% LS error is achieved with 8 modes. It should be noted that, when a small number of modes is used, a thermal discontinuity appears at the element interface located at x = 44 nm or y = 44 nm. As more modes are included, the DG method effectively minimizes the thermal discontinuity.

Table I Eigenenergy (eV) derived from QEM and DNS

QS	POD QS Energy	DNS QS Energy	POD Error (%)
1	0.302586	0.301536	0.347
2	0.312154	0.311371	0.251
3	0.349276	0.348191	0.311
4	0.350074	0.348991	0.310
5	0.354537	0.353177	0.385
6	0.359292	0.358014	0.356
7	0.370174	0.368846	0.359
8	0.374889	0.373563	0.354

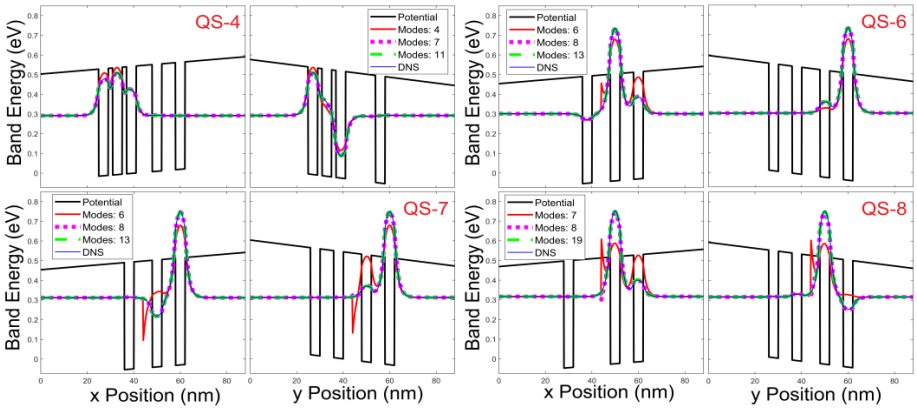


Fig. 4. WFs in QSs 4, 6, 7 and 8 along the paths indicated by the red dashed lines in x and y directions in Fig. 3.

This study demonstrates the effectiveness of the QEM for 2D nanostructure with a reduction in DoF by 4 orders of magnitude. It should be noted that the accuracy of the QEM strongly depends on the quality of the collected data, which is influenced by the training structure, numerical settings, the accuracy of the DNSs used for collecting the WF solution data, etc. Accuracy and efficiency of the QEM are also affected by the thermal discontinuity at the element interfaces, which can be improved by selecting an appropriate range of the penalty factor in the DG method [13,14]. These issues will be further investigated in the extended paper.

Acknowledgements: This work is supported by NSF under Grant Nos. OAC-2118079 and OAC-1852102.

References

[1] A. A. Andronov, E. P. Dodin, D. I. Zinchenko, "Transport in GaAs/$Al_xGa_{1-x}As$ superlattices with narrow minibands: Effects of interminiband tunneling," Semiconductors 43 (2) 228-235, 2009.

[2] Suren A. Tatulian, "From the Wave Equation to Biomolecular Structure and Dynamics," Trends in Biochemical Sciences, 43 (10) 749-751, 2018.

[3] A. J. Cohen, P. Mori-Sánchez, W. Yang, "Insights into Current Limitations of Density Functional Theory," Science, 321 (5890) 792–794, 2008.

[4] H. S. Zhang, et al., "First-Principles Calculations of Quantum Efficiency for Point Defects in Semiconductors: The Example of Yellow Luminance by GaN: CN+ON and GaN:CN, Advanced Optical Materials, 5 (21) 1700404, 2017.

[5] H. Dakhlaoui, J. A. Vinasco, C.A. Duque, "External fields controlling the nonlinear optical properties of quantum cascade laser based on staircase-like quantum wells," Superlattices and Microstructures, vol. 155, 106885, 2021.

[6] G.J. Supran, et al., High-Performance Shortwave-Infrared Light-Emitting Devices Using Core-Shell (PbS-CdS) Colloidal Quantum Dots. Adv. Mater.. 27(8):1437-1442, 2015.

[7] S. A. Veldhuis, et al., "Perovskite Materials for Light-Emitting Diodes and Lasers," Adv. Mater.. 28(32):6804-6834, 2016.

[8] F. P. García de Arquer, et al., "Semiconductor quantum dots: Technological progress and future challenges," Science, vol. 373, no. 6555, eaaz8541, 2021.

[9] M. Veresko, M.C. Cheng, "An Effective Simulation Methodology of Quantum Nanostructures based on Model Order Reduction," *SISPAD* 2021, pp. 64-68, doi: 10.1109/SISPAD54002.2021.9592599.

[10] J. L. Lumley, Atmospheric Turbulence and Wave Propagation, Moscow, Russia, Nauka publisher, 166 (1967).

[11] J. L. Lumley, Stochastic Tools in Turbulence, Academic, New York, 1970; reprint, Dover publisher, 2007.

[12] M.C. Cheng , "Quantum element method for quantum eigenvalue problems derived from projection-based model order reduction" , AIP Advances 10, 115305 (2020).

[13] D. N. Arnold, F. Brezzi, B. Cockburn, D. Marini, Discontinuous Galerkin methods for Elliptic Problems (Lecture Notes in Computational Sci. & Eng.), vol. 11. Berlin, Germany: Springer, 2000, pp. 89–101.

[14] D. N. Arnold, F. Brezzi, B. Cockburn, and L. Donatella Marini, "Unified analysis of discontinuous Galerkin methods for elliptic problems," SIAM J. Numer. Anal., vol. 39, no. 5, pp. 1749–1779, 2002.

[15] L. Sirovich, "Turbulence and the dynamics of coherent structures. I–Coherent structures. II–Symmetries and transformations. III–Dynamics and scaling," Quart. Appl. Math., vol. 45, pp. 561–571 and 573–590, Oct. 1987.

[16] W. Jia, B. T. Helenbrook, and M.-C. Cheng, "Fast Thermal Simulation of FinFET Circuits Based on a Multiblock Reduced-Order Model," IEEE Trans. CAD ICs. Syst, vol. 35, no. 7, pp. 1114–1124, Jul. 2016.

[17] W. Jia, M.C. Cheng, "A Methodology for Thermal Simulation of Interconnects Enabled by Model Reduction with Material Property Variation," J. Comp. Sci., 101665, 2022. Online: doi.org/10.1016/j.jocs.2022.101665.

Vertical GaN Diode BV Maximization through Rapid TCAD Simulation and ML-enabled Surrogate Model

Albert Lu[1], Jordan Marshall[1], Yifan Wang[2], Ming Xiao[2], Yuhao Zhang[2], and Hiu Yung Wong[1*]
[1]Electrical Engineering, San Jose State University, CA, USA
[2]Center for Power Electronics Systems, Virginia Polytechnic Institute and State University, VA, U.S.A
[*]hiuyung.wong@sjsu.edu

Abstract – Motivation and Achievement

GaN is becoming a mainstream semiconductor for RF and power applications, with total market size of over $1 billion [1]. Vertical GaN devices, such as vertical GaN diode, have been widely regarded as one of the most promising candidates for next-generation higher-voltage, high-power applications [2][3]. Due to its wide bandgap (3.4eV), the breakdown field of GaN is 10 times more than that of Si. A GaN diode is expected to have more than 1450 times better Baliga's Figure-of-Merit than Si [2]. However, this is only possible if the diode has a proper edge termination design, such as with guard rings [4] and/or junction termination extension (JTE) [5], which requires a lot of domain expertise and prolonged simulation time due to the huge design space.

In this paper, two methodologies are used to speed up the maximization of the breakdown voltage (BV) of a vertical GaN diode that has a theoretical maximum BV of ~2100V. Firstly, we demonstrated *a 5X faster accurate* simulation method in Technology Computer-Aided-Design (TCAD). This allows us to find *4X more numbers of high BV (>1400V)* designs at a given simulation time. Secondly, a machine learning (ML) model is developed using TCAD-generated data and used as a surrogate model for differential evolution optimization. It can inverse design an *out-of-the-training-range* structure with BV as high as *1805V (85% of the ideal case)* compared to ~1100V designed with human domain expertise.

Simulation Setup

TCAD Sentaurus is used for structure creation and device simulation [6]. Fig. 1 shows an example of the simulation structure. Guard rings are added next to the anode for edge termination. The drift region is 10μm with 10^{16}cm^{-3} n-type doping of silicon. The anode and the guard rings are p-type and doped with 10^{19}cm^{-3} magnesium. Fermi-Dirac statistics, incomplete ionization, high-field mobility saturation, and impact ionization are turned on using the calibrated parameter values in [7]. To maximize the BV, 5 design variables are used. The variables are the space (S) between the guard rings, the width (W) of the guard rings, the depth (D) of the guard rings, the number (N) of the guard rings, and the standard deviation (σ) of the guard ring junction. Fig. 2 shows the BV of an ideal 1D structure of about 2100V which represents the theoretical limit. The current is scaled by assuming the third dimension is 1mm. Using human expertise by experimenting with S, W, D, and N, the highest BV obtained is ~1100V.

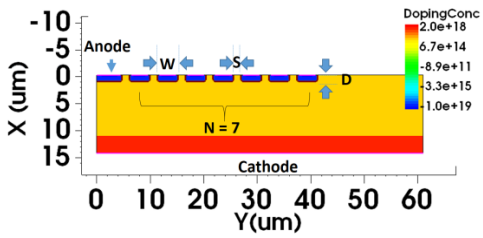


Figure 1: A simulation structure examplar of the vertical GaN diode used in this study. Guard ring number (N) = 7 is used as an example. 4 of the design variables (S, W, D, N) are highlighted. The standard deviation, σ, of the junction gradient is not shown.

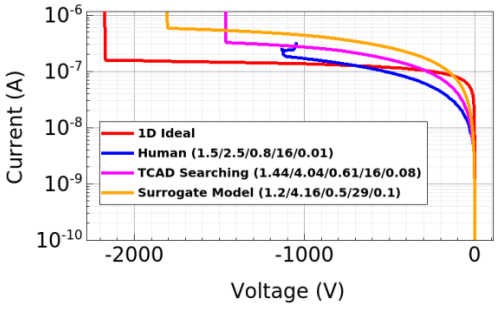


Figure 2: BV curves of selected designs. The third dimension of the 2D structures is set to 1mm. The parentheses contain the values corresponding to S(μm), W(μm), D(μm), N and σ(μm).

BV Maximization through TCAD Searching

We then generate various devices using TCAD by randomly creating structures with $S \in [0.25,5]$, $W \in [0.25,5]$, $D \in [0.01,1]$, $N \in [0,32]$, and $\sigma \in [0.01,0.1]$ all in μm except N which is unitless. 300 structures were simulated and the total simulation time is about 3 days on 30 cores. Only 1 structure is found to have BV > 1400V (Fig. 2) using this random TCAD searching method.

Rapid BV Simulation

It is desirable to find an *accurate* and fast simulation setup to speed up the TCAD searching speed. Such a setup

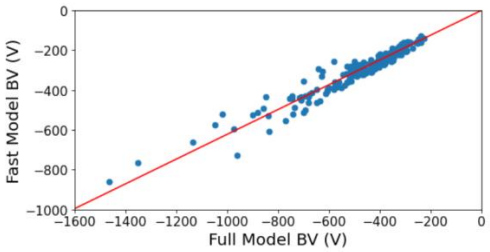


Figure 3: Relationship between the full model BV (with impact ionization) and the fast model BV (without impact ionization and measured at maximum E-field = 3.3MeV/cm). The fitted slope is 0.623.

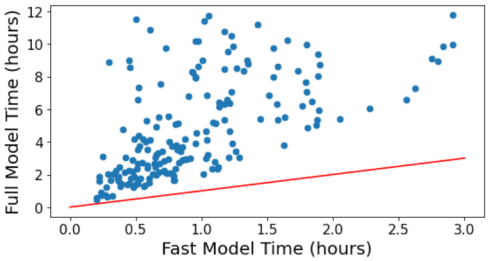


Figure 4: Comparison of the simulation time using the full and fast models. Each point represents one simulated structure. The red line is y = x.

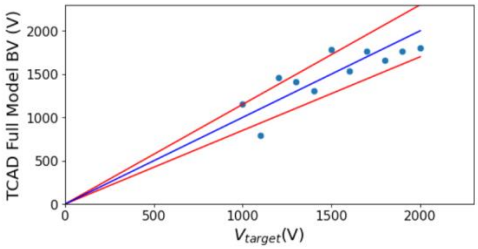


Figure 6: TCAD simulation BV for the inverse designed structures predicted by differential evolution. The red lines are the ±15% boundaries of the expected value (blue line).

can also be used to generate enough data for machine learning (ML). Various BV simulation simplification schemes such as removing incomplete ionization model, not solving hole continuity equations, using ionization integral method, etc. have been tested. Among them, it is found that only removing impact ionization and measuring the peak electric field at 3.3MeV/cm (fast model) provides a significant speedup and accurate solutions. Fig. 3 shows the relationship between the BV obtained using the fast model and the full model which can be modeled as a linear relationship well. Fig. 4 shows the distribution of simulation time of the two models. The speedup can be as much as 24 times and on average, the speedup is 5 times. Moreover, among the 300 simulations, *91% converge using the fast model and only 62% converge with the full model*. The fast model is then used to generate thousands of structures in 3 days and 4 structures are found to have BV > 1400V.

BV Maximization by ML-enabled Surrogate Model

Even though the fast model can speed up the simulation by 5X, it still cannot find a design with higher BV. An ML model is thus developed to correlate the 5 design parameters to a set of full model BV. A neural network (NN) with 1 input layer, 2 hidden layers (each has 50 hidden nodes followed by batch normalization), and 1 output layer is used (Fig. 5). 70%, 15%, and 15% of the data are used for training, testing, and validation, respectively. R^2's are all larger than 92%. The trained machine is then used as a surrogate model for the differential evolution algorithm to design the guard ring based on any given BV. This is achieved by minimizing $\left|f(S,W,D,N,\sigma) - V_{target}\right|$, where f is the output of the ML surrogate model and V_{target} is the target BV. V_{target} from 1000V to 2000V are then fed into the differential evolution algorithm to inverse

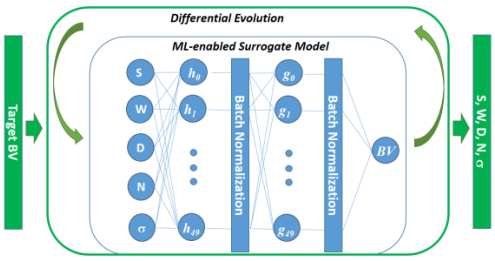


Figure 5: NN used as a surrogate model for differential evolution algorithm to design guard rings for a target BV.

design the diode for the given V_{target}. For example, for V_{target} = 2000V, it is deduced that $S = 1.2, W = 4.16, D = 0.5, N = 29, \sigma = 0.1$ should be used. The corresponding TCAD structures are then constructed and simulated with the full model. Fig. 6 shows that the algorithm can inverse design the device pretty well and most of the BV meets the target within ±15%. Particularly, it can achieve BV = 1805V for V_{target} = 2000V, much higher than its training data (Fig. 2). This is ~85% of the ideal value. Each of the optimizations takes only ~30mins on a laptop.

Conclusions

We proposed a new TCAD setup that has a 5X faster speed and 2X better convergence for GaN diode BV simulation and has a linear correlation to the full model BV. This allows the finding of 4X more designs with high BV (>1400V). To further explore the design with higher BV, a NN is built as the surrogate model and, by using differential evolution, a design with BV as high as 1805V is discovered.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 2134374.

References

[1] Y. Zhang et al., "GaN FinFETs and trigate devices for power and RF applications: review and perspective," Semicond. Sci. Technol., vol. 36, no. 5, p. 054001, Mar. 2021, doi: 10.1088/1361-6641/abde17.

[2] Y. Zhang and T. Palacios, "(Ultra)Wide-Bandgap Vertical Power FinFETs," IEEE Trans. Electron Devices, vol. 67, no. 10, pp. 3960–3971, Oct. 2020, doi: 10.1109/TED.2020.3002880.

[3] T. Oka, "Recent development of vertical GaN power devices," Jpn. J. Appl. Phys., vol. 58, no. SB, p. SB0805, Apr. 2019, doi: 10.7567/1347-4065/ab02e7.

[4] K. Kinoshita, T. Hatakeyama, O. Takikawa, A. Yahata and T. Shinohe, "Guard ring assisted RESURF: a new termination structure providing stable and high breakdown voltage for SiC power devices," Proceedings of the 14th International Symposium on Power Semiconductor Devices and Ics, 2002, pp. 253-256, doi: 10.1109/ISPSD.2002.1016219.

[5] J. Wang et al., "High voltage, high current GaN-on-GaN p-n diodes with partially compensated edge termination," Appl. Phys. Lett., vol. 113, no. 2, p. 023502, Jul. 2018, doi: 10.1063/1.5035267.

[6] Sentaurus™ Device User Guide Version S-2021.06, June 2021.

[7] Sentaurus Technology Template: Simulation of Vertical GaN Devices: Trench-Gate MOSFET and Diodes, Synopsys Inc., 2021.

Hierarchical Modeling for TCAD Simulation of Short-Channel 2D Material-Based FETs

Luca Silvestri[1], Mattias Palsgaard[2], Reto Rhyner[1], Martin Frey[1], Jess Wellendorff[2], Søren Smidstrup[2], Ronald Gull[1] and Karim El Sayed[3]
[1]Synopsys Switzerland LLC, Zurich, Switzerland; [2]Synopsys Denmark ApS, Copenhagen, Denmark; [3]Synopsys Inc., Mountain View, CA, USA

***Abstract*—An integrated hierarchical modeling flow for fast analysis and prototyping of 2D material-based field-effect transistors (FETs) is presented. Advanced transport simulators using *ab initio* atomistic density functional theory (DFT) and continuum effective mass non-equilibrium Green's Functions (NEGF) consistently provide TCAD tools with device material and component parameters and reference curves for physical model selection and calibration. We validate each step of the hierarchical flow, and we show that the simulations performed with the resulting TCAD setup accurately predict 2D-FETs device characteristics.**

I. INTRODUCTION

2D-FETs fabricated with transition metal dichalcogenide (TMD) materials are promising candidates to replace silicon technology at 15 nm channel length and below [1]. However, 2D material technology is still facing several challenges, such as forming good source/drain contacts to TMDs [2] and variability issues [3]. The design of short-channel 2D-FETs requires the investigation of atomic-scale phenomena and the accurate analysis of quantum-mechanical effects. At the same time, the traditional drift-diffusion model (DD) model used in TCAD is infused with more rigorous approaches to maintain accuracy, while still affording the fast turnaround time and flexibility required by modern development in the semiconductor industry [4].

In this work, we present a hierarchical modeling simulation tool, integrated in Sentaurus Materials Workbench tool of QuantumATK software package by Synopsys [5], for fast analysis and prototyping of 2D-FETs. It allows to investigate the impact on device characteristics of different channel materials, number of channel monolayers (ML), source/drain contact material and orientation, device doping and geometry.

II. SIMULATION METHOD

A detailed scheme of the implemented method is shown in Fig.1. As a first step, atomistic DFT calculation of the chosen 2D material is performed with QuantumATK. The material parameters extracted from the results are given as input to Sentaurus Device QTX (indicated with QTX below) [6] to simulate device characteristics using a continuum effective mass (EM) NEGF transport model. 2D ballistic EM NEGF simulations run in a few minutes. The calculated curves are used as reference for the calibration of quantum-corrected carrier density, ballistic mobility and source-to-drain tunneling models in Sentaurus Device [7]. The device component parameters extracted from QuantumATK, that is, the source/drain Schottky barriers, are instead given as input directly to the Schottky contact model in Sentaurus Device. The resulting calibrated model setup can be used for TCAD design of short-channel 2D-FETs.

III. RESULTS AND DISCUSSION

In Fig.2, an example of monolayer (1ML) MoS_2 DFT band structure is reported. The k-space is scanned for minima and the effective mass parameters are automatically extracted by fitting the local band structure.

In order to extract layer thickness and permittivity of the 2D material to be used in continuum models, we implemented a methodology that first extracts the layer thickness from the projected average *ab initio* electron density. The corresponding dielectric constant is found by matching the Hartree Difference Potential (HDP) of a continuum dielectric region with the extracted thickness to *ab initio* model. An example is shown in Fig. 3. The Schottky barrier height (SBH) formed between channel and source/drain contacts is extracted from the calculation of the projected bands of the TMD-metal atomic system (Fig.4).

In Fig. 5, 2D-FETs ballistic Id-Vg curves calculated with EM NEGF and full-band atomistic DFT NEGF are compared. EM NEGF tends to overestimate the current due to narrow bandwidth of TMDs. We implemented a solution that uses the finite bandwidth extracted from DFT calculation as cut-off for the density of states (DOS) in EM NEGF simulations. A good match between EM NEGF and DFT NEGF Id-Vg characteristics is obtained with the finite bandwidth correction, which validates the first step of the model hierarchy.

The Sentaurus Device setup for carrier density used to reproduce QTX characteristics includes a multi-valley description, 2D DOS model and the density gradient (DG) model, to account for quantum confinement in the device cross-section. The DG parameters are calibrated as function of the number of MLs. In Fig. 6, the Sentaurus Device density profile for MoS_2 1ML and 2ML are compared to the corresponding QTX curves. As a result of the accurate density profile calibration, the C-V characteristics of Sentaurus Device and QTX are in good agreement, as illustrated in Fig. 7.

An automatic calibration of the parameters of the ballistic mobility model and the source-drain tunneling model is performed to match the QTX linear and saturation Id-Vg characteristics for different channel lengths down to 5 nm. Some results are reported in Fig. 8, together with density and velocity profiles along the channel. Sentaurus Device well reproduces QTX characteristics, which validates the second step of the hierarchical model.

The hierarchical approach has been tested on different combinations of TMD materials, number of MLs and source/drain metals. The fully automatized parameter extraction and calibration flow runs in a few hours.

Finally, the calibrated Sentaurus Device TCAD setup has been tested by comparing the simulation results with measurements from the literature. An example is illustrated in Fig.9, where we simulated the 10 nm channel 2D-FET from [8].

IV. REFERENCES

[1] K. P. O'Brien at al., IEDM 2021. [2] Terry Y.T. Hung et al., IEDM 2020. [3] Q. Smets et al. IEDM 2020. [4] M. Stettler et., IEDM 2019. [5] QuantumATK version T-2022.03, Synopsys QuantumATK (https://www.synopsys.com/silicon/quantumatk.html). [6] Sentaurus™ Device QTX User Guide, Version T-2022.03. [7] Sentaurus™ Device User Guide, Version T-2022.03. [8] C. D. English et al, IEDM 2016.

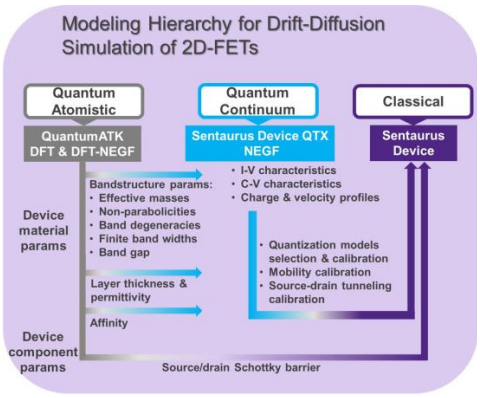


Fig. 1. Schematic of the two-step hierarchical modeling approach implemented in Sentaurus Materials Workbench, with the details of the parameter exchanges.

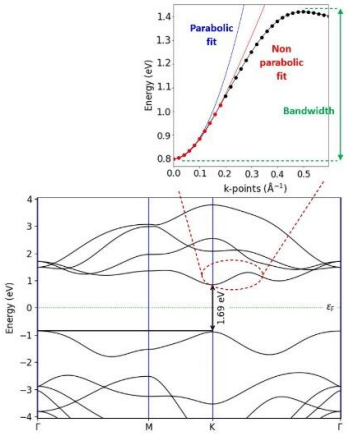


Fig. 2. DFT band structure of monolayer MoS_2. The effective mass parameters are extracted for each valley minima along the two free directions.

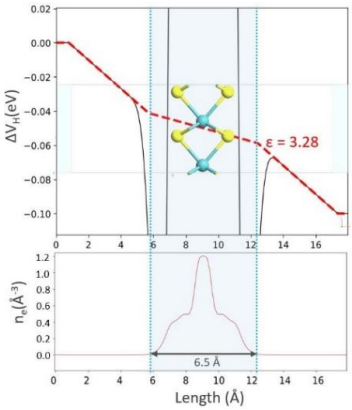
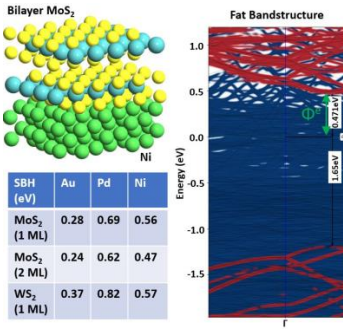


Fig. 3. Top: Hartree Difference Potential of *ab initio* model (black solid line) and continuum model (red dashed line) for MoS_2 1ML. Bottom: *ab initio* electron density for the layer thickness extraction.



SBH (eV)	Au	Pd	Ni
MoS_2 (1 ML)	0.28	0.69	0.56
MoS_2 (2 ML)	0.24	0.62	0.47
WS_2 (1 ML)	0.37	0.82	0.57

Fig. 4. Simulated atomistic 2ML MoS2-Nickel system. The Schottky barrier height (SBH) is extracted from the fat band structure as the distance between the Fermi energy and the lowest energy conduction band with significant projection weight. SBH results for different TMD-metal systems are reported. Metals are (100) oriented.

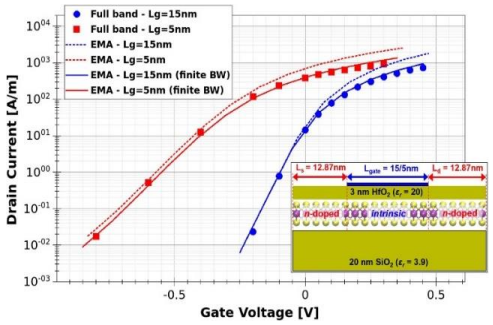


Fig. 5. Id-Vg curves for two different channel lengths calculated with atomistic full-band NEGF (symbols) and continuum EM NEGF (lines) with and without accounting for the finite bandwidth correction. The simulated structure is also shown (for the continuum case the atomistic channel is replaced by a continuum region).

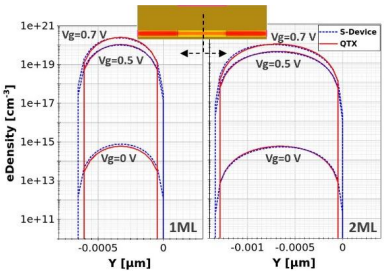


Fig. 6. Comparison between electron density profiles in a device cross-section as a function of the gate voltage calculated with EM NEGF (QTX) and DD (Sentaurus Device) with density gradient.

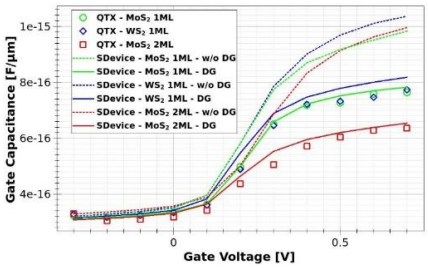


Fig. 7 Cg-Vg curves for three different channels (1ML and 2ML MoS_2 and 1ML WS_2) calculated with EM NEGF (symbols) and DD (lines) with and without accounting for the density gradient correction.

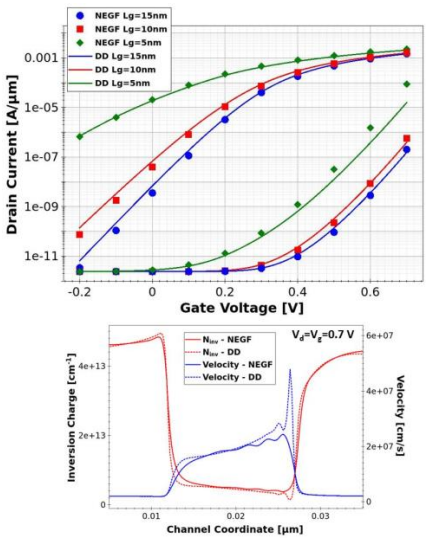
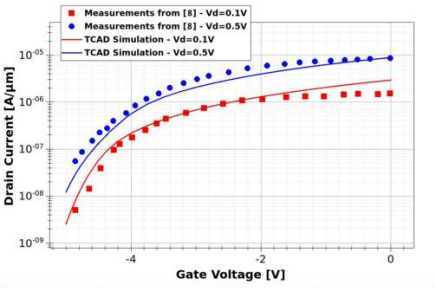


Fig. 8. Top: Id-Vg curves at Vd=0.7 V for three different channel lengths of 1ML WS_2 calculated with EM NEGF (symbols) and DD (lines). Bottom: inversion charge and velocity profiles along the channel.



Quantity	Value	Quantity	Value
2D material	MoS_2	L_{leg}	Full substrate
N° monolayers	1	T_{IL}	0
Device type	Double gate	T_{ox}	5 nm (EOT=5 nm)
Metal contact	Gold (SBH=0.276 eV)	T_{box}	30 nm (EOT=30 nm)
L_{gate}	10 nm	Ch doping	undoped
$L_{contact}$	30 nm	S/D doping	undoped
L_{spacer}	10 nm	μ_{ph}	20 cm²/Vs

Fig. 9. TCAD Id-Vg simulation results compared with the experimental curves from [8]. Bottom: device parameters used in the simulation.

Modeling of SiC Transistor with Counter-doped Channel

Pratik B. Vyas, Ashish Pal, Stephen Weeks, Joshua Holt, Aseem K. Srivastava, Ludovico Megalini, Siddarth Krishnan, Michael Chudzik, El Mehdi Bazizi, and Buvna Ayyagari-Sangamalli
Applied Materials, Santa Clara, USA; Email: pratikb_vyas@amat.com

Abstract: In this paper, we present a modeling framework to simulate the electrical characteristics of SiC MOSFET. Our model also describes the mobility improvement with counter doping in channel. Our analyses show improved drive current and degraded/lower threshold voltage with a counter-doped channel. To this end, we investigate the impact of varying the doping concentrations of the counter-doped region and the underlying p-well for optimum device performance.

Introduction: Field-effect transistors built on 4H Silicon Carbide (4H-SiC) are widely used for high-power applications owing to their high breakdown voltage, thermal conductivity, and ability to form native silicon dioxide [1]. However, performance of these devices is considerably affected by the presence of high density of interface traps (D_{it}) at the gate-oxide/SiC interface [2], which lowers the field-effect mobility and increases on-resistance. In this work, we present a TCAD-based methodology to simulate the electrical characteristics of n-channel SiC MOSFETs and investigate the impact of channel counter-doping [3] as a viable option to counteract the adverse impact of D_{it}. Our results show that the channel counter-doping improves the electron mobility and the device drive-strength, but at the cost of loss/lowering of the threshold voltage (V_T). Finally, we present a comparison of drive current gain and V_T loss for different concentrations of the counter-doped region to achieve the best possible scenario for practical applications.

Device Description: The baseline SiC device-under-study (Fig. 1) has a channel length of 10 µm with an oxide thickness of 50 nm. The source/drain (S/D) regions are implanted with nitrogen (N) with a peak active doping concentration of 1e19 cm^{-3}, while the p-well has a peak aluminum (Al) concentration of 1e18 cm^{-3}. The process modeling is performed in close alignment with experimental setup to facilitate accurate model calibration. A box-shaped counter-doped region with 1e17 cm^{-3} N concentration is added to the channel to study its impact on the device performance.

Simulation: The electrical simulations are performed using drift-diffusion models for electron transport [3]. Due to the high D_{it} concentration at the gate oxide-SiC interface, remote Coulomb scattering with interface charges/trap has a significant impact on the mobility degradation of carriers and has been taken into account. Process flow for the model calibration has been detailed in Fig. 1 (right). As shown in Fig. 2, a uniform D_{it} distribution is considered in the mid-band region and the concentration is calibrated to match the measured sub-threshold slope (SS). Calibration for the above-threshold region is performed using a combination of the mobility model parameters and the D_{it} concentration near the conduction band edge, where an exponential D_{it} distribution is assumed following literature reports [2]. Finally, the D_{it} concentration is fine-tuned to match the conductance peak in the G-V characteristics and the C-V is used as final calibration check. Our analyses show that surface roughness scattering is a dominant mechanism, along with high D_{it} concentration at the channel interface, that degrades electron mobility in SiC channels. We also present two methods to extract the mobility, as shown below, and compare the results:

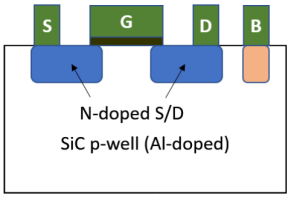
$$g_m\text{-based: } \mu_{FE} = \frac{dI_D}{dV_G}\frac{L}{C_{ox}V_D W} \quad (1) \qquad \text{Weighted average: } \mu_{avg} = \int_0^{s_{max}} \mu(s)n(s)\,ds \Big/ \int_0^{s_{max}} n(s)\,ds \quad (2)$$

Eq. 1 uses the device transfer characteristics to extract mobility, while Eq. 2 calculates a weighted average (weighted against the electron density) of the electron mobility distribution, obtained from TCAD simulations at each bias point, through middle of the channel.

Results and Discussion: The baseline device characteristics in Fig. 4 show close calibration between our SiC model and experimental data. The calibrated D_{it} profile and concentration aligns well with published data [2]. The mobility values obtained from our simulation, highlighted in Fig. 5, are in close alignment with the low channel-mobility values reported in literature for SiC transistors [4][5], validating the significance of our modeling framework. Our analysis shows that scattering with surface roughness, along with the high D_{it} density, at the oxide-SiC interface is the dominant factor causing mobility degradation in these SiC devices. Counter-doping the channel leads to drive current and mobility improvement. However, our results show that an important factor for consideration is the extent of depletion of the counter-doped (CD) region by the underlying p-well. A fully depleted CD region, as a consequence of relatively high p-well doping (similar p-well concentration as our baseline device), leads to a 70% improvement in the drive current (Fig. 6). A combined increase in electron mobility (Fig. 7(a)) as well as a wider conduction path (Fig. 7(b)) accounts for the higher drive current. However, high peaks in mobility above 200 $cm^2/V.s$, as reported in literature [2][4] for CD channels, are achieved only if the CD region is not fully depleted by the underlying p-well, leading to the formation of a buried channel. This is achieved in Fig. 8(a) by reducing the peak p-well doping by an order of magnitude and the formation of a buried channel is observed in Fig. 8(b). However, the resulting 8x increase in drive current is accompanied by a significant loss in the threshold voltage (V_T), as reported in Fig 6, essentially creating an "Always On" device. On the other hand, the V_T loss associated with a fully depleted CD channel can be recovered by tuning the doping profiles and transistor design. Fig. 9 shows that by varying the concentration of the CD region, we can achieve an optimized device design with good drive-strength combined with acceptable V_T loss. It can be seen from Fig. 9(a) that, for as low as 0.5V V_T loss, a decent 10% improvement in drive current can be obtained by a lower 2e16 cm^{-3} CD concentration. Fig. 9(b) also shows that, even at low CD concentrations, the device with low p-well doping and buried channel exhibits significant V_T loss, rendering it unfavorable for practical applications.

Conclusion: We present a TCAD model, validated with experimental data and published literature, to simulate the electrical behavior of the SiC planar MOSFET. Our results show that channel counter-doping indeed improves electron mobility by screening the negative impact of D_{it} and provides a wider conduction path, leading to improved device drive-strength at the expense of V_T loss. The doping concentration of the CD channel can be varied to achieve the most optimum scenario with good current gain for an acceptable V_T loss. We also highlight that the p-well doping level should be high enough so that the CD region is almost fully depleted. Alternatively, a buried channel formation can lead to very high mobility and current gains but at an unrecoverable loss of gate control.

References: [1] L. Kimoto et al., Jpn. J. Appl. Phys. 54, 040103)2015); [2] T. Doi et al, Jpn. J. Appl. Phys. 61, 021007 (2022); [3] Synopsys TCAD Suite, Version 2021.06; [4] K. Ariyoshi et al., Appl. Phys. Lett. 106, 103506 (2015); [5] D. Okamoto et al., IEEE Elec. Dev. Lett. 31, 710 (2010).



Parameter	Dimension
Channel length	10 µm
Gate oxide	50 nm
Peak Channel Doping	~1e18 cm^{-3}
Peak S/D Doping	1e19 cm^{-3}

Fig. 1: Schematic of baseline SiC nMOS (left) and important device dimensions (right)

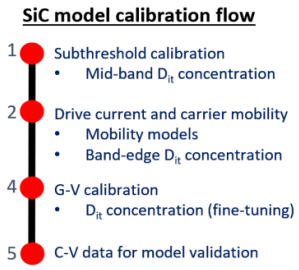


Fig. 2: Flow-chart of model calibration for SiC nMOS

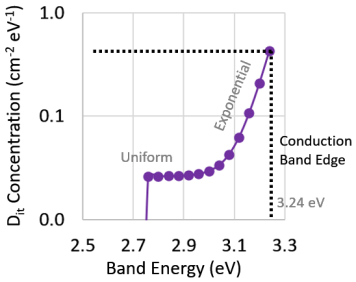


Fig. 3: Dit profile used to calibrate model. Peak Dit concentration aligns with published data [3].

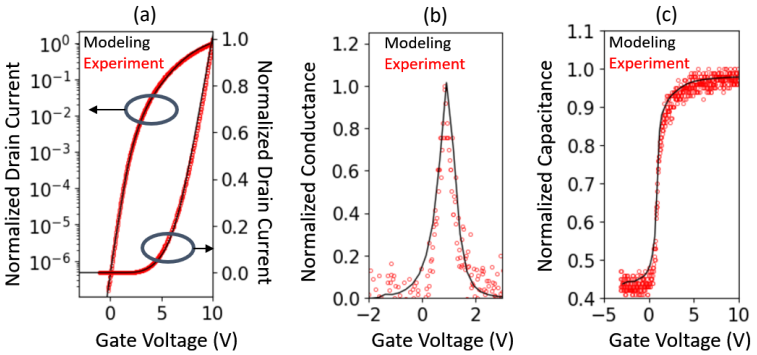


Fig. 4: Final Model calibration vs. experimental data for **(a)** transfer characteristics, **(b)** Conductance vs. Vg, and **(c)** C-V. The transfer characteristics were used to calibrate mid-band and band-edge D_{it}. The G-V was used to fine-tune the final D_{it} concentration. Final model calibration check using C-V

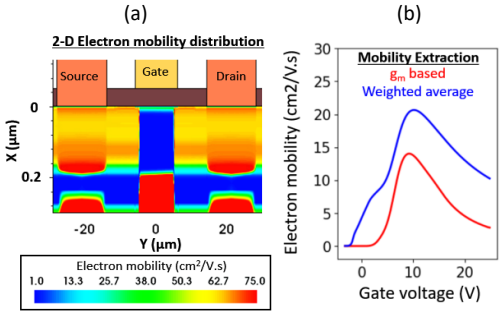
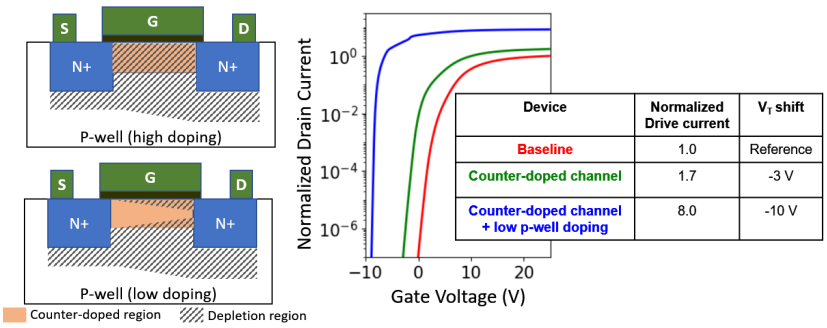


Fig. 5: (a) Electron mobility distribution, and **(b)** Mobility vs Vg, in baseline device, along with comparison between different mobility extraction methods.



Device	Normalized Drive current	V_T shift
Baseline	1.0	Reference
Counter-doped channel	1.7	-3 V
Counter-doped channel + low p-well doping	8.0	-10 V

Fig. 6: Channel counter-doped (CD) with 1e17 cm-3 N till a depth of 200 nm. I-V comparison between baseline device (no CD region), devices with partially depleted CD region, and fully-depleted CD regions. CD channel provides current gain but at the expense of VT loss/lowering. Gate control loss with partially depleted CD region due to very low VT

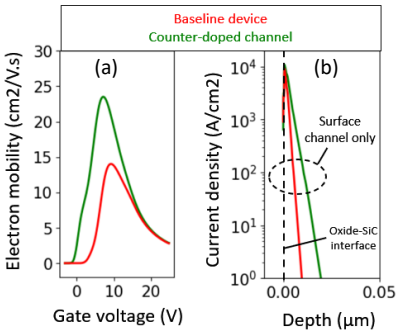


Fig. 7: (a) Channel counter-doping (CD) impact on electron mobility for fully-depleted CD region. **(b)** Width of conduction channel increases but no buried channel formation

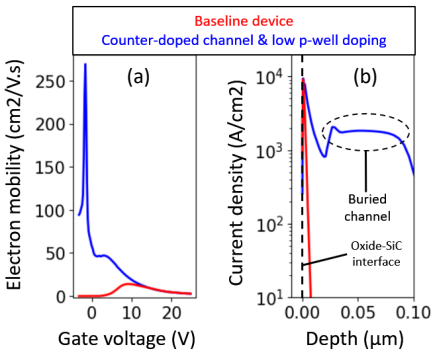


Fig. 8: (a) Channel counter-doping (CD) impact on electron mobility for partially depleted (low p-well doping) CD region. **(b)** Increase in channel width and formation of buried channel. High peaks in mobility only observed when buried channel is formed

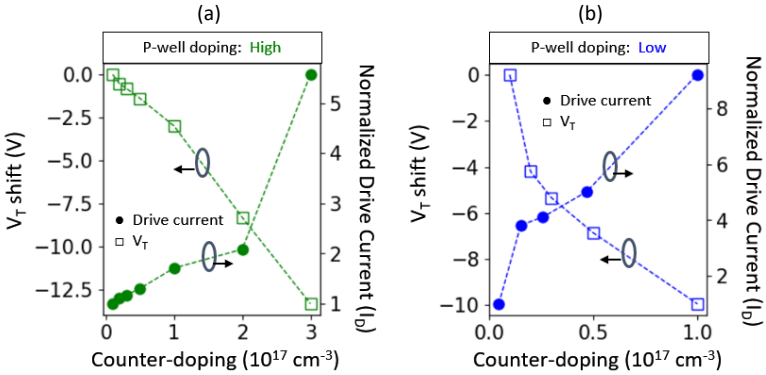


Fig. 9: Varying CD concentration for achieving optimum conditions with **(a)** fully-depleted CD region and **(b)** partially-depleted CD region. Drive current improvement with minimal or acceptable V_T loss can be obtained with the fully-depleted channel. V_T loss in the partially-depleted CD channel case is irrecoverable even after lowering the counter-doping concentration, rendering the design unfavorable for practical application.

Towards a DFT-based layered model for TCAD simulations of MoS_2

L. Donetti, C. Marquez, C. Navarro, C. Medina-Bailon, J. L. Padilla, C. Sampedro, F. Gamiz
Departamento de Electronica and CITIC, Universidad de Granada, Granada, Spain.

***Abstract*—In this work, we employ the results of atomistic DFT calculation to extract useful parameters for the simulation of few-layers MoS_2 structures with traditional TCAD tools. In particular we focus on the charge distribution, which allows us to obtain a layered model for the dielectric constant, and on the effective densities of states in the conduction and valence bands, taking into account the full 2D density of states.**

I. Introduction

The layered structure of 2D semiconductors such as MoS_2 presents several challenges for their accurate simulation employing standard TCAD tools. Apart from the need of thickness-specific material parameters, descriptions based on bulk parameters cannot reproduce accurately the layered spatial charge distribution and the 2D density of states. Therefore, even if the properties of MoS_2 are being successfully studied through *ab-initio* atomistic methods, it is not straightforward to include this knowledge in classical TCAD tools. In this paper, we model the layered structure of MoS_2 through a stack of semiconductor layers separated by insulating layers representing the Van der Waals (VdW) gaps [1] with material parameters extracted directly from Density Functional Theory (DFT) calculation. In particular, we obtain a layered dielectric model through the analysis of charge distribution in mono- and few-layer structures with an applied out-of-plane electric field, and extract layer-dependent N_c and N_v values from the calculation of the full Density Of States (DOS).

II. DFT calculations

DFT calculations are performed with QuantumATK (version S-2021.06) [2], employing its LCAO calculator with GGA exchange/correlation, PBE functional, Grimme DFT-D2 VdW correction. Spin-orbit interaction is included to obtain a better description of the band structure near the valence band maximum. We consider few-layer MoS_2 structures with a number of layers, N, between 1 and 10. We compute the density of states for each structure in equilibrium, to obtain the DOS. Then, we apply an electric field perpendicular to the semiconductor layers by forcing a potential difference between the boundaries of the simulation cell. For the biased structures, we compute the induced electron density difference Δn, and the electrostatic potential difference ΔV, taking in-plane averages as a function of the out-of-plane position z.

III. Layered model

Some examples of the results of the calculations with an applied bias are shown in figure 1. The behavior of the total

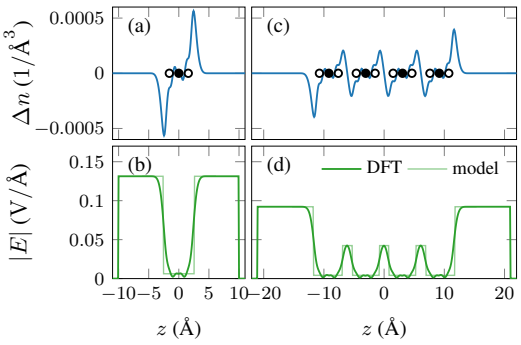


Fig. 1. Excess electron density Δn (a,c) and electric field E (b,d) for monolayer (a–b) and 4-layer (c–d) MoS_2 with 2 V bias. In (a) and (c), closed and open circles represent the z positions of Mo and S atoms, respectively.

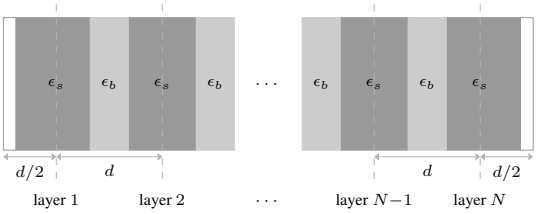


Fig. 2. Layered model for MoS_2. The interlayer distance, $d = 6.115$ Å, is equal to half of bulk MoS_2 c lattice constant.

charge density and the dipole moment density in each layer as a function of the external electric field suggests a layered dielectric model, where MoS_2 layers are separated by VdW gaps with a different dielectric constant, as the one shown in figure 2. By fitting the DFT results, the two dielectric constants and the thickness of the different layers can be extracted [3]. The results suggest that the thickness of the external layers is different from that of the internal ones and also that an empty gap is needed to "fill" the thickness of the multilayer structure, computed as a multiple of the interlayer distance [3].

It is not straightforward to model the DOS of 2D materials through the usual 3D expressions involving the effective mass because of several reasons. First of all, it must be taken into account that the position of the conduction band maxima and valence band minima in k space vary with the number of layers, N, and also the corresponding curvature which defines the effective mass can change. Then, even assuming that the

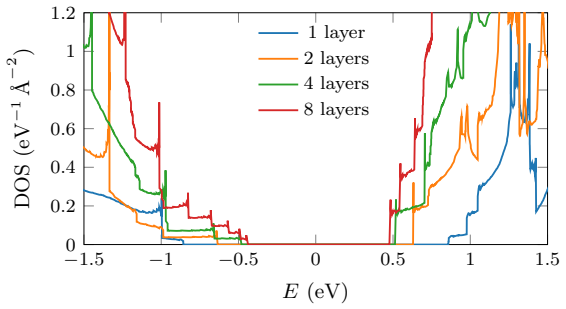


Fig. 3. DOS of different MoS_2 structures with different number of layers.

material parameters can vary with N, the DOS of 2D bands presents a functional dependence on energy which is different from the one for 3D materials. As an example, in figure 3 we show the DOS of mono- and some few- layers structure. In each case, the DOS is approximately constant near the extrema of each band and discrete steps can be observed when the minimum (or maximum) of a different band is reached. Therefore, the usual expressions of the effective densities of states N_c and N_v proportional to the $m_{\mathrm{DOS}}^{3/2}$ (where m_{DOS} is the DOS effective mass) do not hold for 2D materials. To overcome this problem, we compute the effective DOS in the conduction band, $N_{c,2\mathrm{D}}$, with the following expression, which takes into account the whole DOS:

$$N_{c,2D} = \int_{E_c}^{\infty} g(E) \exp\left((E - E_c)/kT\right) dE \qquad (1)$$

where E_c is the conduction band minumum, $g(E)$ is the DOS, k is boltzmann constant, T the absolute temperature and non-degenerate statistics is assumed. The results of equation (1) and an analogous one for $N_{v,2\mathrm{D}}$ are shown in figure 4(a). The jump between mono-layer and bi-layer is due to the difference in the position of band extrema (for $N = 1$ the band gap is direct at the K point and it becomes indirect for $N \geq 2$ [4]) and the increase for larger N is caused by the fact that the following bands get closer in energy to the first one. However, this increase is much smaller than the volume increase due to the larger number of layers: the corresponding 3D parameters obtained taking into account the total semiconductor thickness of the different layered structures shows a pronounced decrease as N increases, as shown in figure 4(b).

To check the implementation of the layered model, we simulated with Sentaurus TCAD [5] the capacitance of metal/oxide/semiconductor structures composed by a 1 nm SiO_2 layer and MoS_2 with different number of layers (inset in figure 5) with the parameters extracted in the previous sections. In figure 5, we show the simulated capacitance compared to the one obtained for similar structures where MoS_2 layers are substituted by a uniform material with averaged parameters. At zero bias, the results obtained with the layered and uniform

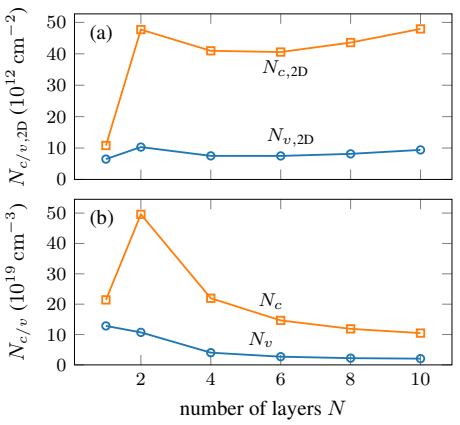


Fig. 4. 2D (a) and 3D (b) effective DOS of MoS_2 structures as a function of the number of layers N.

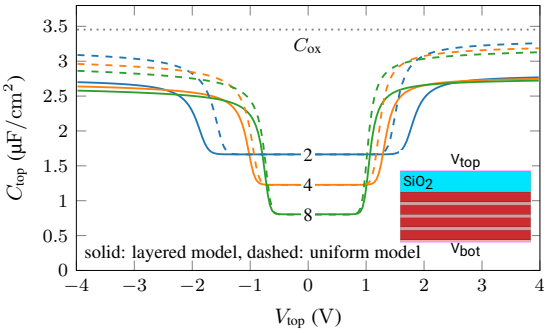


Fig. 5. Capacitance obtained with TCAD simulations of the structure depicted in the inset, for the layered (solid lines) and uniform (dashed lines) models for different number of layers N. The oxide capacitance C_{ox} is shown as a reference.

models are the same, while the results differ for large positive and negative values of the applied voltage.

IV. Conclusions

Employing DFT simulations, we extracted dielectric constants and effective DOS in the conduction and valence bands for the implementation of a layered MoS_2 model to be employed in TCAD simulations.

References

[1] G. Mirabelli, P. K. Hurley *et al.*, "Physics-based modelling of MoS_2: the layered structure concept," *Semiconductor Science and Technology*, vol. 34, no. 5, p. 055015, Apr 2019.
[2] S. Smidstrup, T. Markussen *et al.*, "QuantumATK: an integrated platform of electronic and atomic-scale modelling tools," *Journal of Physics: Condensed Matter*, vol. 32, no. 1, p. 015901, Oct 2019.
[3] L. Donetti, C. Navarro *et al.*, "Dft-based layered dielectric model of few-layer mos$_2$," *Solid-State Electronics*, vol. accepted for publication, 2022.
[4] A. Kuc, N. Zibouche *et al.*, "Influence of quantum confinement on the electronic structure of the transition metal sulfide TS_2," *Physical Review B*, vol. 83, no. 24, Jun 2011.
[5] *Synopsys Sentaurus Device User Guide (T-2022.03)*, 2022.

A Simulation Physics-Guided Neural Network for Predicting Semiconductor Structure with Few Experimental Data

QHwan Kim[1], Sunghee Lee[1], Ami Ma[1], Jaeyoon Kim[1], Hyeon-Kyun Noh[1], Kyu Baik Chang[1], Wooyoung Cheon[1], Shinwook Yi[1], Jaehoon Jeong[1], BongSeok Kim[2], Young-Seok Kim[2], Dae Sin Kim[1]

[1]Computational Science and Engineering Team, [2]Memory Metrology and Inspection Technology Team, Samsung Electronics Co., Hwasung, Gyeonggi, 18448, Korea, Email:qhwan.kim@samsung.com

INTRODUCTION

The accurate estimation of spectroscopic measurement - semiconductor CD relationships with the machine learning is crucial as the semiconductor structure becomes complex and length scale shrinks. However, in an industrial field, the experimental data is few including noise because measurement of the CD, which requires electron microscopy such as transmission electron microscopy (TEM), is time- and cost-consuming process, while the spectrum can be relatively easily measured from optical measurement such as ellipsometry [1]. In this paper, we propose an end-to-end two-step PGNN algorithm, which uses simulation-based model to train experimental data. In the first step, we train broad simulation data and construct solid physical model describing general spectrum-CD relationship. In the second step, we train the experimental model whose loss function is guided by physics obtained from the first step. The proposed algorithms can avoid overfitting induced by small-sample and provide higher prediction accuracy beyond other benchmark algorithms.

METHODS

Table 1 shows the details of six in-line datasets. Each dataset corresponds to the different CD from sequential fabrication process and contains pairs of ellipsometry spectrum (x) and CD (y). Train and test datasets contain experimental measurements, which are obtained from different lots. We additionally use simulation data, which calculates spectrum from 3D semiconductor virtual structure via rigorous coupled-wave analysis.

Fig. 1 shows the schematic of the proposed PGNN model. The model is composed with two sub-models, $f(\cdot)$ and $g(\cdot)$, which train the simulation and experimental data respectively. Two sub-model structures are same and composed with 1D convolutional layers, flatten layer and fully-connected layers. The 1D convolutional layers are used to obtain spectrum representation, and the fully-connected layers act as a classifier.

At first, $f(\cdot)$ is trained with the simulation data. Then weights of $f(\cdot)$ 1D convolutional layers are transferred to the model $g(\cdot)$ and are frozen. At last, $g(\cdot)$ is trained with physics-guided loss function [3] as follows

$$L_{g(\cdot)} = L_{\text{MSE}} + L_{\text{PHYS}}$$
$$= \text{MSE}(y_{\text{pred}}, y) + \sum_{f(x_i)<f(x_j)} \text{ReLU}\left(g(x_i) - g(x_j)\right)$$

where $\text{ReLU}(\cdot)$ denotes rectified linear unit function. The L_{MSE} represents typical empirical loss while the L_{PHYS} represents regularization loss guided by physics from the $f(\cdot)$. L_{PHYS} directly guides to regularize $g(\cdot)$ to follow spectrum-CD relationship defined from $f(\cdot)$.

For comparison, we compare our model against three baselines: PLS, Ridge, and NN, which use only empirical loss L_{MSE} for training $g(\cdot)$. PLS and Ridge convert fully-connected layer of $g(\cdot)$ to the linear models, partial least squares and ridge regressor respectively.

RESULTS AND CONCOLUSION

Fig. 2 visualizes true and predicted spectrum-CD relationships of DATA6. Spectrum features are reduced by principal component analysis (PCA) and CDs are represented by colors. Fig. 2 shows that the simulation data covers broad spectrum range while the experiment data can cover only narrow space. PGNN algorithm can predict spectrum-CD relationships broadly. However, the ridge algorithm without physics regularizer L_{PHYS} only can predict narrow regions around experimental data.

Table 2 lists the RMSE of each algorithm on the train and the test dataset. Table 2 shows that the PGNN outperforms for all test datasets except DATA2, which shows similar RMSE. PLS, Ridge and NN perform better on train datasets, which indicates they are not generalized for unseen spectrum data and suffered from overfitting. The average RMSE of PGNN is 2.22 nm, which is improved 51 % relative with baseline algorithms. Fig. 3 shows the prediction results of PGNN and Ridge, where X and Y axes indicate the predicted and true CDs, respectively. As shown in the Fig .3, PGNN shows significant improvements in the prediction accuracies.

We test the robustness of the PGNN model by varying the data quality. We add the noise to the CDs of train dataset. The noise is controlled by the $\text{Noise Scale} \times N(0, 1^2)$, where Noise Scale is defined to be the percentage ratio of CD average and $N(0, 1^2)$ is standard normal distribution. Fig. 4 shows the RMSEs of all algorithms as a function of an additive noise level. Only PGNN maintains relatively good performance regardless of varying data quality. Fig. 5 shows the ΔRMSE, which is defined by the difference of maximum and minimum RMSE with the label noise data. Because PGNN is robust against external noise, it shows smallest ΔRMSE value. Our work indicate that the PGNN can be used safely in the in-line semiconductor manufacturing process where the internal and external fabrication condition varies continuously due to its accuracy and robustness.

REFERENCES

[1] Ellipsometry Academy: Advance your ellipsometry knowledge and skills Spectroscopic.

[2] R. S. Rai, *Prog. Cryst. Growth Charact. Mater. 55*, 2009, pp. 63-97.

[3] A. Daw, *arXiv:1710.11431v3*, 2017.

	DATA1	DATA2	DATA3	DATA4	DATA5	DATA6
Train#	12	12	293	15	9	9
Avg.	213.89	257.23	28.77	130.39	29.12	84.01
Std.	11.07	13.7	3.11	10.19	2.33	7.9
Test #	18	18	123	6	10	14
Avg.	222.32	254.76	26.83	130.59	33.42	79.29
Std.	9.56	10.9	2.16	7.17	1.38	4.47
Sim. #	1906	1906	1988	1988	1988	1988
Avg.	221.30	254.67	30.19	103.40	28.99	64.19
Std.	11.55	18.22	11.89	27.91	4.21	25.67

Table 1. Details of prepared six benchmark datasets.

	Train Set RMSE (nm)			
	PLS	**Ridge**	**NN**	**PGNN**
DATA1	**0.583**	0.591	0.654	2.183
DATA2	1.13	**0.369**	1.085	2.585
DATA3	0.987	0.983	**0.982**	1.155
DATA4	1.03	1.031	**1.029**	1.39
DATA5	0.478	0.501	0.487	**0.269**
DATA6	0.407	0.407	0.408	**0.335**
	Test Set RMSE (nm)			
	PLS	**Ridge**	**NN**	**PGNN**
DATA1	5.911	9.105	6.610	**3.073**
DATA2	3.442	**3.141**	3.162	3.155
DATA3	3.329	3.340	3.500	**1.847**
DATA4	3.099	3.095	2.954	**2.461**
DATA5	4.245	2.384	4.302	**1.647**
DATA6	6.241	6.292	6.325	**1.138**

Table 2. RMSE performance comparison of each algorithm on the train and test datasets. Boldface represents the best performance model.

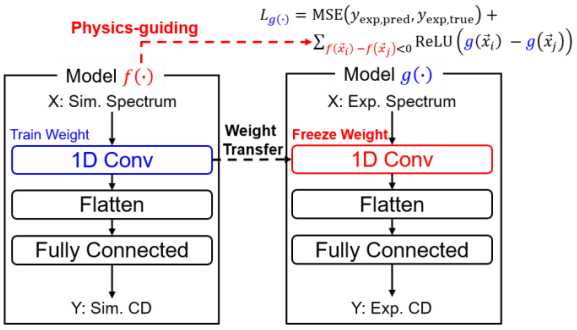


Figure 1. An overview of the proposed PGNN architecture schematic.

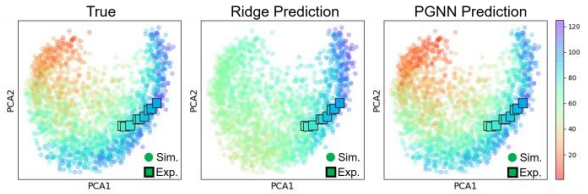


Figure 2. True and predicted spectrum - CD relationships of DATA6. Spectrums are decomposed by PCA algorithms and CDs are represented by colors.

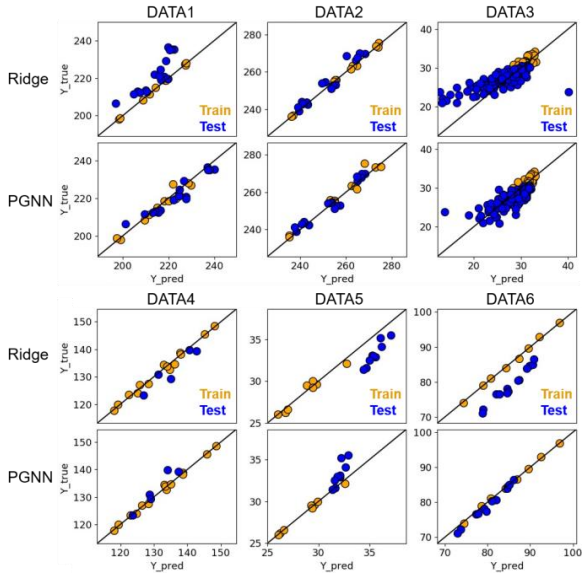


Figure 3. Prediction results of Ridge and PGNN on the train (yellow dot) and test (blue) datasets. The X and Y axes represent the predicted and true CD values, respectively.

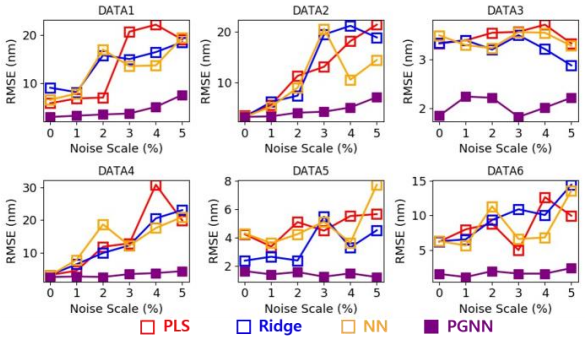


Figure 4. RMSE of algorithms as a function of the Gaussian label noise scale added to the train data

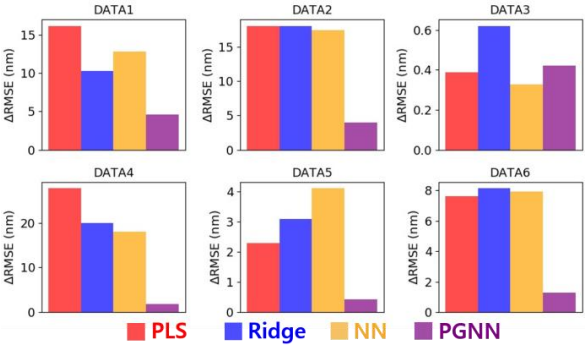


Figure 5. ΔRMSE, which is defined by the difference of maximum and minimum RMSE with the label noise data shown in Fig. 4, of algorithms.

Building Robust Machine Learning Force Fields by Composite Gaussian Approximation Potentials

Diego Milardovich[1], Dominic Waldhoer[1], Markus Jech[1], Al-Moatasem Bellah El-Sayed[1], and Tibor Grasser[1]
[1]Institute for Microelectronics, Technische Universität Wien, Gußhausstraße 27–29, 1040 Vienna, Austria
E-mail: [milardovich | waldhoer | jech | el-sayed | grasser]@iue.tuwien.ac.at

***Abstract*—The use of machine learning (ML) interatomic potentials as part of device simulations has received a lot of interest in recent years, motivated by their high accuracy at low computational costs. However, these potentials have a tendency to overfit, which threatens their transferability. This work proposes a systematic solution to this problem, by augmenting these potentials with a set of simpler machine learning models. The versatility of the proposed solution is demonstrated by developing a machine learning force field for amorphous silicon dioxide (a-SiO_2), exhibiting a significant improvement in transferability, at only a moderate increase in computational costs.**

I. Introduction

With the increasing necessity of including atomistic calculations into simulation workflows for nanoelectronic devices, the interest in machine learning (ML) interatomic potentials as an alternative to more expensive *ab initio* methods like density functional theory (DFT) has arisen. This trend is further aided by readily available software packages giving easy access to ML potentials [1–4]. However, the development of transferable ML interatomic potentials remains a topic of active research in this field. One of the main obstacles which weakens the transferability of these potentials is overfitting, an undesired statistical phenomena on which ML models present a high accuracy on training data, but perform poorly in real applications.

In this work, a systematic solution to mitigate the effects of overfitting is proposed by combining several independent instances of the established Gaussian approximation potential (GAP) [5], a type of kernel-based model successfully used in previous interatomic potential developments [6–8]. This model assumes that similar local atomic configurations give similar contributions to the total potential energy. It estimates the potential energy of a given atomic configuration by comparing it to those provided in a training dataset. This model and its typical implementation in ML interatomic potentials are depicted in Fig. 1. Instead of building an interatomic potential as a single highly-complex ML model, we propose to also include simpler models, which are less likely to suffer from overfitting. We will demonstrate our proposed framework by building a potential for amorphous silicon dioxide (a-SiO_2).

II. Methodology

In a traditional ML potential, the total potential energy of a given atomic system is approximated by a sum of local energy contributions from every atom, computed by a **single ML model** [9],

$$E_{\text{total}} = \sum_{i}^{\text{Atoms}} E_i(\boldsymbol{d}_i),$$

where $\boldsymbol{d}_i$ is a local descriptor for the environment of the i-th atom and $\boldsymbol{E}_i$ is its local contribution to the total potential energy. We propose to build a potential which computes these local energy contributions as a **composite set of multiple ML models**, i.e.,

$$E_i = \sum_{j}^{\text{Models}} E_i^j(\boldsymbol{d}_i^j),$$

where j runs over all ML models composing the final potential. Each of these models can be based on different local descriptors, allowing for varying degrees of complexity. In this work we build an interatomic potential for a-SiO_2 composed of a simple auxiliary potential, which represents the basic physics by pairwise short-range interactions, as shown in Fig. 1(c), and a more complex main model responsible for giving accurate results for the system of interest,

$$E_i = E_i^{\text{main}}(\boldsymbol{d}_i^{\text{main}}) + E_i^{\text{aux}}(\boldsymbol{d}_i^{\text{aux}})$$

In our implementation, both potentials are realized as GAPs. The auxiliary potential uses a simple two-body descriptor [3], whereas the main potential employs the more sophisticated *smooth overlap of atomic positions* (SOAP) [10] descriptor. The auxiliary potential was trained on a set of 3000 diatomic configurations (Si-Si, Si-O and O-O) with interatomic distances ranging from 0.50 to 5.00 Å. The training dataset for the main GAP was created by running molecular dynamics (MD), according to the melt-and-quench technique [11] within the LAMMPS engine [12]. A defect-free 216-atoms SiO_2 system was melted at 5000 K and subsequently quenched to 300 K, using the classical force-field ReaxFF [13] with a time-step of 0.25 ps. The process is depicted in Fig. 2. The resulting trajectory was sequentially sub-sampled to a training dataset of 1500 atomic configurations, for which the energies were calculated with DFT, using the PBE functional [14] in the CP2K software package [15]. The main potential was trained on the residual between the DFT energies and the predictions of the auxiliary model for the configurations in this dataset. The training was performed using the software package QUIPPY [3]. The two-body descriptor for the short-range auxiliary GAP was defined using $\mathrm{r}_{\text{cut}}^{\text{SiSi}} = 1.60$ Å, $\mathrm{r}_{\text{cut}}^{\text{SiO}} = 1.10$ Å and $\mathrm{r}_{\text{cut}}^{\text{OO}} = 0.80$ Å, while the SOAP descriptor for the main GAP was constructed with $\mathrm{n}_{\text{max}} = 6$, $\mathrm{l}_{\text{max}} = 6$ and $\mathrm{r}_{\text{cut}}^{\text{SOAP}} = 4.0$ Å.

III. Results and Discussion

Following the traditional approach, a single GAP was trained on the dataset used to train the main ML model of our proposed composite potential. This single potential and our composite potential were evaluated by computing the potential energies in a testing dataset of 1000 a-SiO_2 structures, yielding a similar energy accuracy of roughly 5 meV/Atom, as shown in Fig. 3. However, when running an MD with these potentials, as shown in Fig. 4, the structures produced by the traditional model suffer from unphysical atomic cluster formations, as is apparent from the radial density functions (RDFs). In contrast, our composite potential shows excellent agreement with the DFT reference structure. The reason for this is that training datasets built by running MD rarely contain short-range interatomic interactions, as they are high in energy and therefore unlikely to be present in the MD trajectory. This prevents the traditional ML potential from learning that short interatomic distances correspond to high energies, resulting in the unphysical clustering seen in the MD results. Since this information is explicitly included in the auxiliary potential, our approach is much more robust while also retaining a high accuracy, with only little computational overhead (roughly 5%).

To validate the transferability of our composite ML potential, the vibrational density of states (VDOS) for one of the resulting a-SiO_2 structures was computed and validated against DFT results, as shown in Fig. 5(a). The results are remarkably similar, considering that the ML potential was only trained on energies but not atomic forces. We further tested our potential by performing MD calculations for systems 10-times the size of the training dataset structures, as shown in Fig. 5(b), where no unphysical behavior is present.

IV. Conclusions

ML models are a powerful tool for developing highly accurate interatomic potentials. However, when underlying physical mechanisms are not explicitly taken into account, their transferability is in question. In our proposed approach, the ML potential is augmented by a set of simpler ML models. Results show that this significantly reduces the unphysical behavior when facing unexplored atomic environments, at a moderate increase in the computational costs.

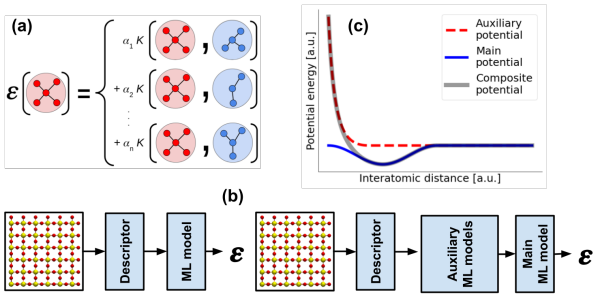


Fig. 1: (a) Schematic of a GAP, a ML model which computes the potential energy of a given atomic configuration as the weighted sum over the similarity measurements between the given atomic configuration and those in a training dataset. The similarities are measured by a kernel. Training this model is equivalent to finding the optimal weight values $(\alpha_1, \alpha_2, ..., \alpha_n)$. (b) Workflows for a traditional potential built with this ML model (left) and our proposed composite interatomic potential (right). (c) Schematic of the proposed composite potential, augmenting the main potential with an auxiliary potential.

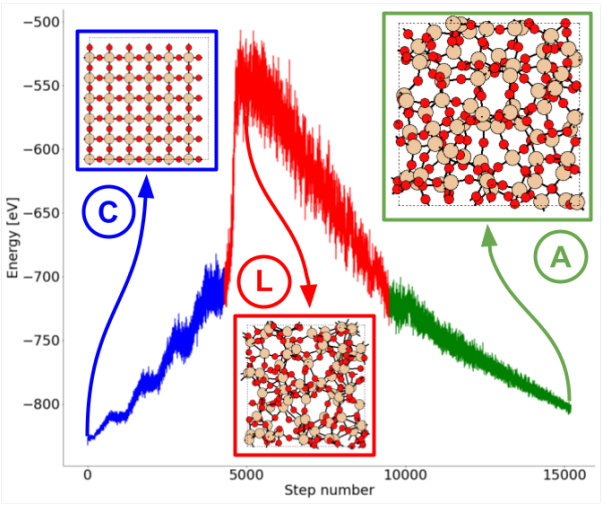


Fig. 2: Initial dataset for training and testing the *main* ML potential, created using the melt-and-quench technique and the ReaxFF force field. The MD begins with a crystalline SiO_2 structure, which is melted up to 5000 K and thereafter quenched back to 300 K. Once the dataset is created, the energies are recalculated on a subset using DFT. Blue: Crystalline phase (C). Red: Liquid phase (L). Green: Amorphous phase (A).

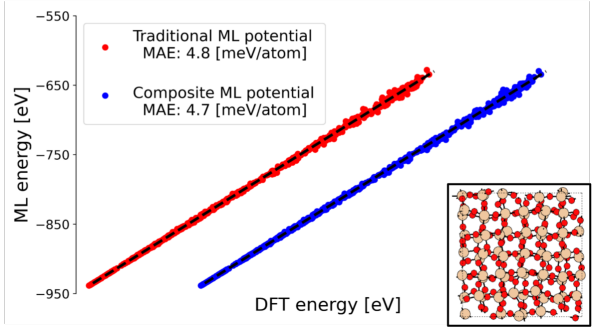


Fig. 3: Testing the potential energy accuracy for the traditional (red) and the composite (blue) ML potentials against DFT. The results show a high accuracy for both potentials and virtually no difference between them. An example of the atomic structures used for testing is shown in the bottom-right sub-panel.

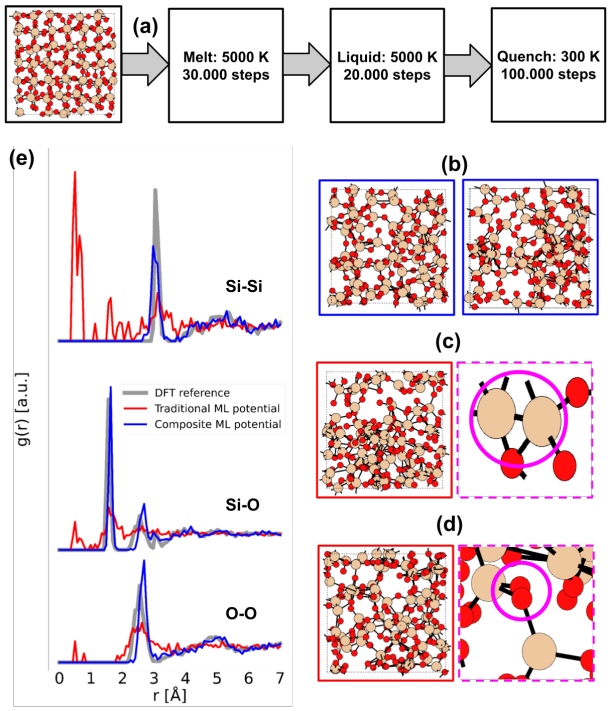


Fig. 4: Panel comparing our proposed framework results with those of a traditional ML interatomic potential, built as a single GAP. Both potentials were used to run the same MD, as specified in (a). Two of the resulting structures when using our proposed approach are presented in (b). Resulting structures from using the traditional ML potential are presented in (c) and (d), together with examples for unphysical behavior found in them. The RDFs for one of the structures built with our approach and one built with the traditional ML potential are presented in (e), together with the reference from the training dataset. As it can be seen, the composite potential performs much better when compared against the DFT reference.

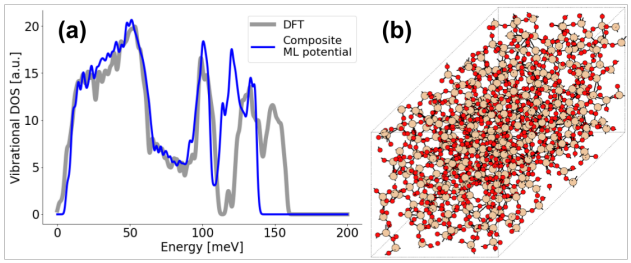


Fig. 5: Transferability tests for our proposed composite ML potential. (a) Computing the VDOS and comparing it to DFT results. The good match is remarkable, considering that the potential was not trained on forces. (b) Using the potential to create an a-SiO_2 structure 10 times the size of those used in the training dataset. No unphysical clustering is present.

References

[1] A. Khorshidi et at. *Comput. Phys. Commun.*, 207:310–324, 2016.
[2] H. Wang et al. *Comput. Phys. Commun.*, 228:178–184, 2018.
[3] J. R. Kermode. *J. Phys. Condens. Matter*, 32:305901, 2020.
[4] R. Lot et al. *Comput. Phys. Commun.*, 256:107402, 2020.
[5] A. P. Bartók et al. *Phy. Rev. Lett.*, 104:136403, 2010.
[6] G. Sivaraman et al. *npj Comput. Mater.*, 6:104, 2020.
[7] A. P. Bartók et al. *Phys. Rev. X*, 8:041048, 2018.
[8] V. L. Deringer et al. *Phys. Rev. B*, 95:094203, 2017.
[9] J. Behler et al. *Phys. Rev. Lett.*, 98:146401, 2007.
[10] A. P. Bartók et al. *Phys. Rev. B*, 87:184115, 2013.
[11] Al-Moatasem El-Sayed et al. *Microelectron. Eng.*, 209:68–71, 2013.
[12] S. Plimpton. *J. Comput. Phys.*, 117:1–19, 1995.
[13] J. C. Fogarty et al. *J. Chem. Phys.*, 132:174704, 2010.
[14] J. P. Perdew et al. *Phys. Rev. Lett.*, 77:3865, 1996.
[15] J. VandeVondele et al. *Comput. Phys. Commun.*, 167:103–128, 2005.

Surrogate models for device design using sample efficient Deep Learning

Rutu Patel[1], Nihar R. Mohapatra[1], and Ravi S. Hegde[1]
[1]Indian Institute of Technology Gandhinagar, Gujarat, India, patel.rutu@iitgn.ac.in

Abstract

Generation of training dataset for machine learning-based device design algorithms, is expensive. To take care of this, sampling techniques using active learning are proposed in this work. Their efficiency is demonstrated through a DNN-based LDMOSFET off-state breakdown voltage ($BV_{DS,off}$) and specific on resistance (R_{sp}) predictor. Results show a ~50% reduction in dataset size without compromising the accuracy. Particularly, I-QBC and DI-GS work best with ~1.87% ENPE.

Introduction

TCAD tools, which solve the physical equations at set mesh points, have been developed and used over the years to reduce the device design time and cost. However, the computational time for devices with a large number of mesh points is more (e.g. for high electric field simulations in the prediction of $BV_{DS,off}$ of LDMOSFETs). To tackle such bottlenecks, data driven surrogate models which mimic TCAD are being developed [1-3]. Popularly, DNNs trained using supervised learning (labeled samples from the input feature space) are used. The prediction accuracy of these models improves as the dataset size increases. As generating a large dataset is not always feasible, developing surrogate models for complex devices with a large number of input features is also difficult.

Active learning, a technique of choosing the best samples from a pool such that the accuracy of the surrogate model improves, was proposed by Dongrui Wu [4]. Three criteria were suggested to choose the samples- Informativeness (I), Representativeness (R) and Diversity (D). In this work, 8 different sampling techniques which promise to reduce the training dataset size are developed: I-GS (Informative Greedy Sampling), I-QBC (Informative Query by Committee), DI-GS (Diverse I-GS), DI-QBC (Diverse I-QBC), R-GS (Representative GS), R-QBC (Representative QBC), DR-GS (Diverse R-GS) and DR-QBC (Diverse R-QBC).

Methodology

Fig.1a shows an LDMOSFET structure. With the prior experience of TCAD based design approach, 7 design parameters and their ranges (mentioned in Table I) are chosen as input feature space. These parameters along with the output variables ($BV_{DS,off}$ & R_{sp}) form the surrogate model shown in Fig. 1b. Euclidean Norm of the Prediction Error (ENPE) is chosen as the accuracy FoM and is plotted for a separately generated test dataset of 300 Latin Hypercube Samples (LHS).

$$P.E.(\%) = 100 \times |(Actual - Predicted)/Actual|$$

$$ENPE\ (\%) = \sqrt{P.E.BV_{DS,off}^2 + P.E.R_{sp}^2}$$

To optimize the configuration of DNN based predictor (to achieve balance between under fitting and over fitting), experiments were performed using training dataset of 100 LHS (Fig. 1c). As evident, the optimized predictor which provides the minimum ENPE, has 3 hidden layers with 64 neurons each. Fig. 1d shows that with the same predictor, ENPE further reduces by increasing the training dataset size and is ~1.86% for 1100 LHS. Further, to reduce the dataset size with the same accuracy, the sample efficient algorithm as depicted in Fig. 2, is developed. Combining the methods of (I) choosing the base samples (B), (II) calculating the Euclidean distance of remaining samples, and (III) choosing the most or the least distant additional samples (m), we have proposed 8 techniques.

Results and discussion

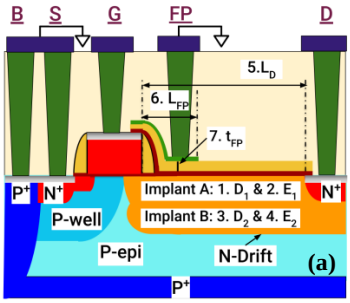
The ENPE for all these techniques are extracted by performing the experiments on 3 types of datasets with different N (pool size), B and m (Fig. 3). ENPE reduces as total number of selected samples increases and it saturates as the number approaches N. The saturation ENPE reduces as N increases (Fig 3a-c and Fig 3d-e). The dataset with N=900, B=200 and m=75 is found to be sufficient to achieve baseline ENPE of ~1.87%. Specifically, the I-QBC and DI-GS techniques work best and provides the same ENPE with only 500 training samples. Fig. 4 shows the R_{sp}-$BV_{DS,off}$ tradeoff for the surrogate model trained by 500 samples chosen using the DI-GS technique. Actual and predicted values for the test dataset are in good agreement.

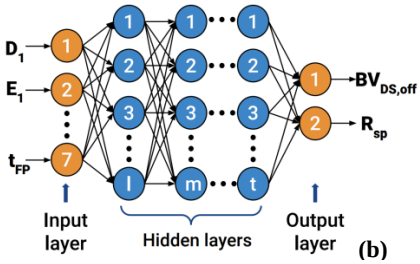
Conclusion

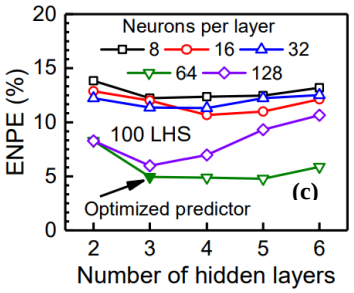
50% reduction in the training dataset size is achieved by the I-QBC and DI-GS sampling techniques. These benefits can be leveraged by using surrogate models for complex devices and in inverse design.

References

[1] Jing Chen et al, IEEE TED, Sept 2021 [2] Hiro Gangi et al, ISPSD, May 2021 [3] Kashyap Mehta et al, IEEE Access, Aug. 2020 [4] Dongrui Wu, IEEE TNNLS, May 2019.







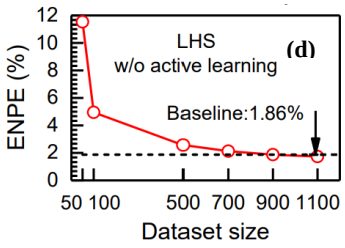


Fig. 1 a) Schematic of the LDMOSFET structure showing the device design parameters (Table I) that affect $BV_{DS,off}$ and R_{sp}. b) The DNN based predictor surrogate model. c) ENPE vs. number of hidden layers with different number of neurons per layer plot to choose optimum predictor. d) ENPE reduces by increasing training dataset size, for the chosen optimized predictor. The baseline accuracy is 1.86% of ENPE for 1100 LHS samples w/o active learning.

Table I

Design parameter	Range
Implant A Dose (D_1)	0.8-1.2x10^{12} cm^{-2}
Implant A Energy (E_1)	160-220 keV
Implant B Dose (D_2)	0.8-1.2x10^{12} cm^{-2}
Implant B Energy (E_2)	260-320 keV
Drift length (L_D)	1.5-9 µm
Field plate length (L_{FP})	0.5-4 µm
Field plate dielectric thickness (t_{FP})	0.5-0.8 µm
Design constraints: $E_1<E_2$, $L_{FP}<L_D/2$	

Proposed 8 efficient sampling techniques (I x II x III):
I-GS, I-QBC, DI-GS, DI-QBC,
R-GS, R-QBC, DR-GS, DR-QBC

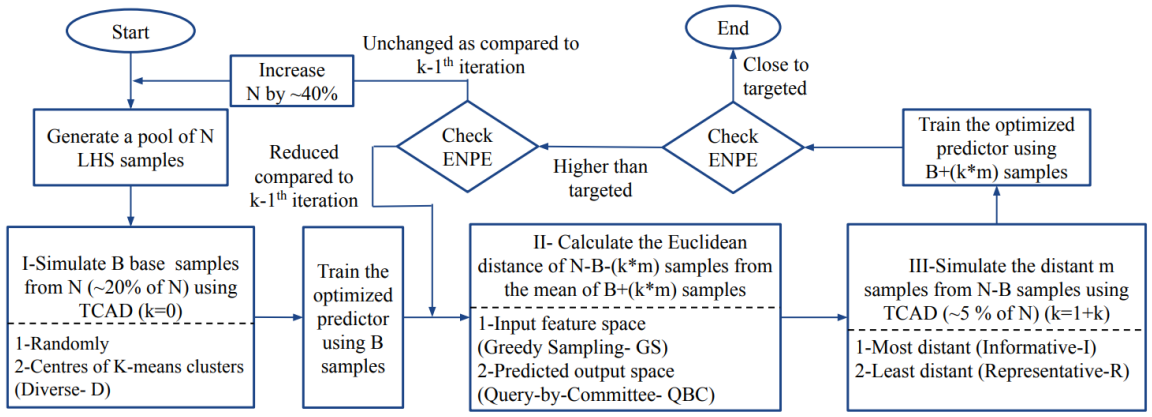


Fig. 2 Flow chart depicting the efficient sampling algorithm. 8 different sampling techniques are proposed, considering the possible the methods of I) choosing B, II) calculating Euclidean distance and III) choosing the distant m samples.

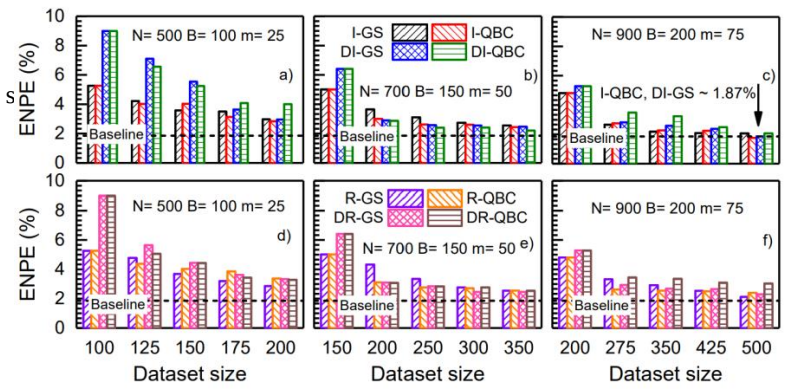


Fig. 3 ENPE of the predictors trained using 8 active learning based sampling techniques for 3 different datasets. N, B and m increases from a) to c) and from d) to e). I-QBC and DI-GS sampling techniques achieve ~1.87% ENPE with total 500 samples (N=900, B=200 and m=75) as compared to 1100 LHS samples w/o active learning.

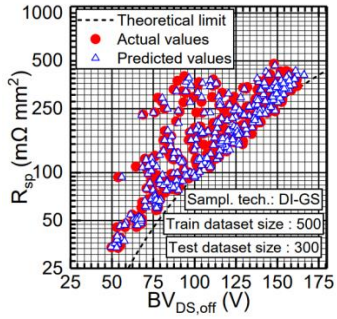


Fig. 4 The R_{sp} vs. $BV_{DS,off}$ plot of actual and predicted values of the test dataset. The most efficient DI-GS sampling technique with total 500 samples is used to train the predictor.

TCAD Augmented Generative Adversarial Network for Optimizing a Chip-level Size Mask Layout Design in the HARC Etching Process

Hyoungcheol Kwon [1, 3, *], Hwiwon Seo [3], Hyunsuk Huh [2], Felipe Iza [1], Dongyean Oh [3], Sung Kye Park [3], and Seonyong Cha [3]
[1]The Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, Leicestershire, United Kingdom
[2]Department of Mechanical Engineering, POSTECH, Pohang-si, Gyeongsangbuk-do, 37673, Republic of Korea
[3]R&D Devision, SK Hynix Inc., Icheon-si, Gyeonggi-do, 17336, Republic of Korea
*email: h.kwon2@lboro.ac.uk

***Abstract*—Even though multiscale technology computer-aided design (TCAD) methodology is suitable for effectively predicting etching processes and optimizing recipes (RCP), it is highly time-consuming. This article demonstrates that our deep learning platform called TCAD-augmented Generative Adversarial Network can reduce the computational load by 2,600,000 times. This platform opens up new applications such as hot spot detection and mask layout optimization in a chip-level area of 3D NAND fabrication.**

I. Introduction

In 3D NAND fabrication, new structures and processes must be developed to maintain the evolutionary path for z-scaling, which inevitably leads to a significant increase in structural complexity and cost. As such, vertical etching of plug holes and word lines is a crucial success factor in enhancing manufacturability [1-7]. Although the shape of the layout is faithfully transferred on the wafer, the difference between mask layout and cross-sectional completed etch profile is significant, as shown in Fig. 1 (a). Thus, with optical proximity correction (OPC), advanced and reliable etch proximity correction (EPC) is required during the layout design stage [8]. Recently, 3D feature profile simulation using technology computer-aided design (TCAD) has been widely applied to predict etching processes and optimize RCP, including mask layout design [9, 10]. As shown in Fig 1 (b), the proposal of mask layout design can be searched to meet process targets using a thousand trials of simulation within a few days. However, this TCAD based EPC is highly time-consuming; thus, simulations over an entire chip is not feasible. Instead, it is more feasible to use TCAD to generate finite data sets to train deep learning platforms [16]. In other words, using an advanced deep learning methodology can dramatically reduce the simulation time [17-19]. In this article, for the first time, we address TCAD-augmented Generative Adversarial Network (TaGAN) for chip-level hot spot detection and EPC in the HARC etching.

II. TCAD Modelling

The multiscale TCAD methodology was used to investigate plasma etching profiles, and it comprises three parts. In the first stage, the plasma characteristics such as the fluxes of ions and neutrals to the wafer and ion energy and angle distributions were obtained from the reactor level 0D global modelling [11-13]. The dependence of HARC etching profiles on the plasma characteristics and other conditions can be analyzed in the second stage [14]. The last is post-processing work related to the time and spatial analysis of the profiles. Details of the multiscale TCAD methodology have been described in our previous work [15].

III. TCAD-Augmented Deep Learning Modelling

Figure 2 shows the dataset (500×500 nm) used to train our deep learning, i.e., TaGAN. Each etching process needs its mask layout database, and the domain of an entire chip (~mm) layout could be divided into a hundred nm, i.e., cell-level size, as shown in Fig. 3. Then, we chose different 200 cases with cell-level sizes from the chip-level layout database, and corresponding computational 3D etching profiles were obtained using TCAD. Flip and rotation were used to train data. In turn, we generated more than 1600 pairs of the dataset for training. Figure 4 shows the detailed structure of the implemented TaGAN model. The overall learning structure is similar to the GAN structure. However, our model has additional image regression loss, and an image is put into the Generator to create a new fake image instead of a noise input. The structure of UNet, mainly used for segmentation, is used as the Generator after modification to enable the image-to-image regression. The Discriminator constructed five convolution blocks using the structure of DenseNet. DenseNet has a complex structure because a dense block makes many skip connections, but this structure has good feature extraction performance and classification performance [16-25].

IV. Results And Discussion

Figure 5 shows input mask images and corresponding etch profile images among the test sets according to the various models. Etching profile predictions of the UNet and TaGAN model are similar to the TCAD-only results compared with the VAE predictions [26]. Moreover, UNet and TaGAN prediction does not overfit training data even for small datasets because model architecture has encoder and decoder structures with skip connections. Both UNet and TaGAN models predict well, but the expressive ability, e.g., line edge roughness, is higher in the TaGAN. Quantitatively similarity of each deep learning model with TCAD-only can be compared using various metrics indicators as shown in Table I [27-29]. All indicators denote that TaGAN predicts the most similar to TCAD-only calculation. Figure 6 shows the computational time using TCAD-only and TaGAN for the 20 test mask input data. It takes 60 to 80 hours through TCAD-only, but the TaGAN models only take about 0.1 seconds with the same accuracy.

V. Summary and Conclusions

Although TCAD methodology has excellent interpretability and explainability for etching process analysis, it is highly time-consuming because many physical and chemical reactions are incorporated into the models. For this reason, we demonstrated the deep learning

platform called TaGAN and its detailed architecture. Incorporating deep learning such as GAN into TCAD shows a promising solution for dramatically reducing the simulation time. For the first time, more than a hundred datasets of 3D etching profiles for various layout designs are generated using a well-calibrated physical model for the training learning platform. TaGAN demonstrated that it could reduce the computational load by 2,600,000 times with the same accuracy as TCAD-only, even when the actual experimental data are scarce. Its high accuracy of similarity with TCAD-only is verified using the various similarity metrics. In turn, the simulations for the weak-point detection, such as not opening and bridging during HARC etch processes in an entire chip, can be done within a few days. Our TaGAN platform opens up new applications such as hot spot detection and masks layout optimization in a chip-level area of 3D NAND fabrication.

Table I. Similarity metrics between images generated by TCAD-only and by deep learning models such as VAE, UNET, and TaGAN.

Method	RMSE ↓	PSNR ↑	SSIM ↑	MSSSIM ↑	FID ↓
VAE	54.52	13.57	0.6618	0.7232	305.98
UNet	6.37	32.65	0.9667	0.9842	158.07
TaGAN	**5.51**	**34.03**	**0.9689**	**0.9880**	**97.04**

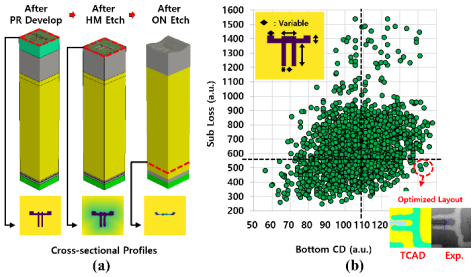


Fig. 1. (a) Simplified fabrication steps for high stack etchings and their cross-sectional etch profiles at the bottom. (b) Optimization example of mask layout design using a thousand simulation trials.

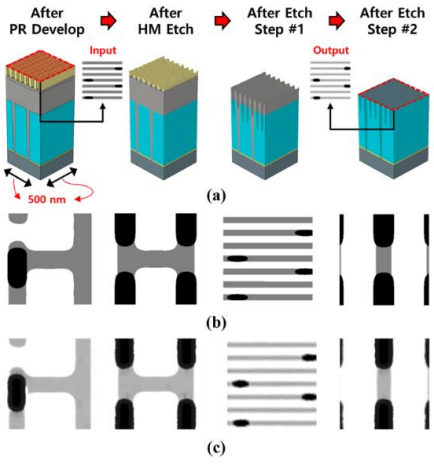


Fig. 2. (a) Process steps for high stack etchings and dataset extraction position. Dataset examples of cross-sectional (b) mask layout images at the top of the structure (input for the train) and (c) their corresponding completed etching profiles (output for the train)

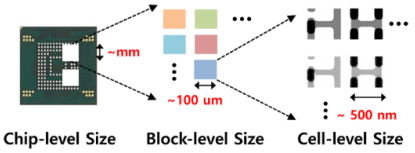


Fig. 3. The relative size scale of chip-, block-, and cell-level layouts of 3D NAND Memory.

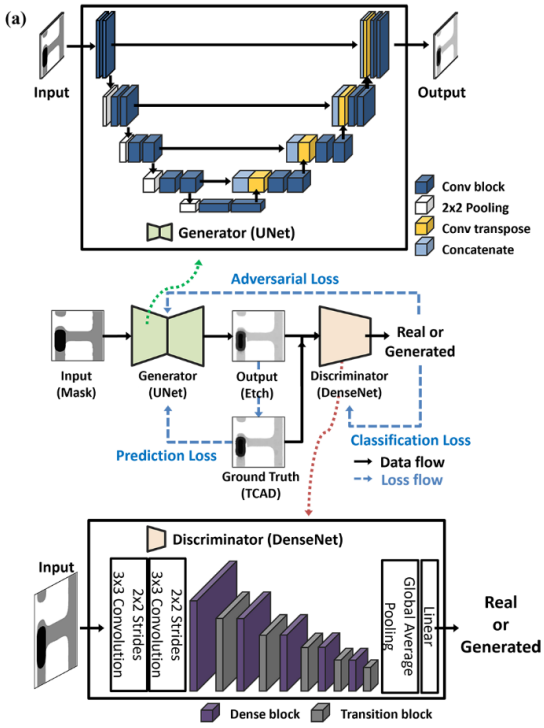


Fig. 4. The schematic diagrams of TCAD-augmented generative adversarial network for etching profile prediction. The UNet was used as the Generator, and DenseNet was used as the Discriminator of the TaGAN.

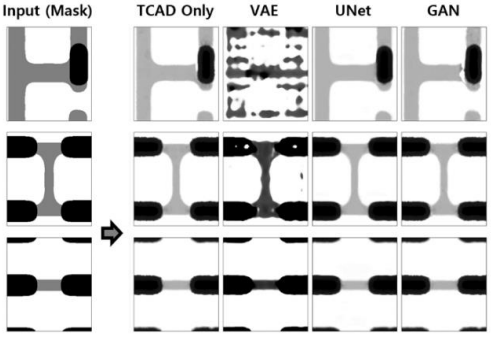


Fig. 5. Cross-sectional mask layout (input) at the top of the structure and their completed etching profile prediction (output) at the bottom of the structure for various methodologies such as TCAD-Only, VAE, UNet, and TaGAN.

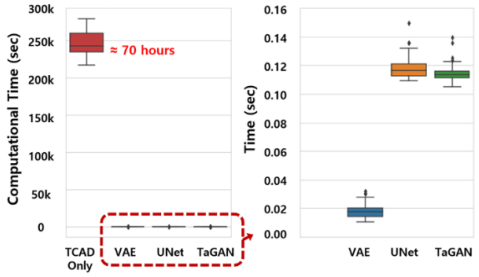


Fig. 6. Computational time comparison with different methodology in the case of 20 different etching profile predictions of 300×300 size (500×500 nm)

References

[1] S. H. Lee, *IEEE Int. Electron Devices Meeting* (IEDM), pp. 1.1.1-1.1.8 (2016).

[2] S. K. Park, *IEEE Int. Memory Workshop* (IMW), pp. 1-4 (2015).

[3] R. Katsumata, M. Kito, Y. Fukuzumi, M. Kido, H. Tanaka, Y. Komori, M. Ishiduki, J. Matsunami, T. Fujiwara, Y. Nagata, L. Zhang, Y. Iwata, R. Kirisawa, H. Aochi, and A. Nitayama, *Symposium on VLSI Technology*, pp. 136-137 (2009).

[4] J. Jang, H. Kim, W. Cho, H. Cho, J. Kim, S. Shim, Y. Jang, J. Jeong, B. Son, D. Kim, K. Kim, J. Shim, J. Lim, K. Kim, S. Yi, J. Lim, D. Chung, H. Moon, S. Hwang, J. Lee, Y. Son, U. Chung, and W. Lee, *Symposium on VLSI Technology*, pp. 192-193 (2009).

[5] E. S. Choi and S. K. Park, *IEEE Int. Electron Devices Meeting* (IEDM), pp. 9.4.1-9.4.4 (2012).

[6] K. Parat and C. Dennison, *IEEE Int. Electron Devices Meeting* (IEDM), pp. 3.3.1-3.3.4 (2015).

[7] C. Cho, K. You, S. Kim, Y. Lee, J. Lee, and S. You, Materials **14**, 5036 (2021).

[8] P. Parashar, C. Akbar, T. S. Rawat, S. Pratik, R. Butola, S. H. Chen, Y. Chang, S. Nuannimnoi, and A. S. Lin, IEEE Photonics Journal, **11**, 2800215 (2019)

[9] K SPEED User Guide, KW Tech (2021).

[10] Sentaurus Topography 3D User Guide, Synopsys (2021).

[11] D. C. Kwon, D. H. Yu, H. C. Kwon, Y. H. Im, and H. C. Lee, Phys. Plasmas **27,** 073507 (2020).

[12] H. S. You, Y. G. Yook, W. S. Chang, J. H. Park, M. J. Oh, D. C. Kwon, J. S. Yoon, D. H. Yu, H. C. Kwon, S. K. Park, and Y. H. Im, J. Phys. D:Appl. Phys. **53**, 385207 (2020).

[13] W. S. Chang, Y. G. Yook, H. S. You, J. H. Park, D. C. Kwon, M. Y. Song, J. S. Yoon, D. W. Kim, S. J. You, D. H. Yu, H. C. Kwon, S. K. Park, and Y. H. Im, Applied Surface Science **515**, 145975 (2020).

[14] Y. G. Yook, H. S. You, J. H. Park, W. S. Chang, D. C Kwon, J. S. Yoon, K. H. Yoon, S. S. Shin, D. H. Yu, and Y. H. Im, J. Phys. D:Appl. Phys. (2022). in press

[15] H. C. Kwon, “Effect of heavy inert ion strikes on cell density dependent profile distortion during HARC etching process”, unpublished

[16] H. Dhillon, K. Mehta, M. Xiao, B. Wang, Y. Zhang, and H. Y. Wong, IEEE Transactions on Electron Devices **68**, pp. 5498-5503 (2021).

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", Proc. NIPS, pp. 2672-2680 (2014).

[18] L. Dinh, D. Krueger, S. Bengio, "Nice: Non-linear independent components estimation", arXiv 1410.8516 (2014).

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", *Proc. Int. Conf. Medical Image Comput. Comput.-Assisted Intervention*, pp. 234-241 (2015).

[20] C. -W. Teo, K. L. Low, V. Narang and A. V. -Y. Thean, *International Conference on Simulation of Semiconductor Processes and Devices* (SISPAD), pp. 1-4 (2019).

[21] S. Saini, K. Lata, G.R. Sinha, VLSI and Hardware Implementations Using Modern Machine Learning Methods (CRC Press, Boca Raton, 2021).

[22] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis", arXiv:2111.14791 (2021).

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks", *Proceedings of the 34th International Conference on Machine Learning* (PMLR) **70**, pp. 214-223, arXiv:1701.07875 (2017).

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANS", *Proc. Advances Neural Information Processing Systems Conf.*, arXiv:1704.00028 (2017).

[25] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2261-2269 (2017).

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", arXiv:1312.6114 (2013).

[27] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM", *Proc. IEEE International Conference on Pattern Recognition* (ICPR), pp. 2366-2369 (2010).

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium", *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, pp. 6626-6637 (2017).

[29] A. Borji, "Pros and cons of gan evaluation measures", Computer Vision and Image Understanding **179**, pp. 41-65 (2019).

Mono-material TMD-based heterostructures for nanoelectronics applications

Farzan Gity
Low-dimensional Heterogeneous Integration
Nanoelectronics Materials and Devices, Micro/Nano Systems Centre
Tyndall National Institute, University College Cork, Ireland

Transition metal dichalcogenide (TMD) based heterostructures can be formed by interfacing two different TMDs through vertical stacking in which these TMDs are interacting through weak van der Waals (vdW) forces, to circumvent the conventional lattice-mismatch problem. Heterostructures of two dissimilar TMDs can be used in device structures with atomically sharp and clean interfaces. TMD-based heterostructures allow for exploring new physics and device architectures by utilising different strategies, such as combining different TMDs, various crystallographic alignment and stacking sequence [1, 2]. Combining monolayers of different TMDs in vertical and lateral geometry allows for manipulating the electrical and optical properties of the heterostructure-based devices.

Molybdenum disulfide (MoS_2) is the most studied TMD with a wide variety of applications such as field-effect transistors (FETs), light harvesting devices, chemical sensors, photocatalysts, and flexible electronics [3-8]. Bulk MoS_2 is semiconducting with an indirect bandgap of 1.2 eV, whereas monolayer MoS_2 is a direct gap material with a bandgap of 1.8 eV, where the conduction band minimum and the valence band maximum are located at the K-point [9, 10]. Tungsten disulfide (WS_2) possesses similar crystal structure as of MoS_2 [11, 12] where monolayer WS_2 is also direct semiconductor with bandgap of 1.97 eV. Both electron and hole effective masses of WS_2 are smaller than that of MoS_2. AA' stacking of monolayer MoS_2 and WS_2 creates an indirect bandgap material with larger contribution of WS_2 valance and MoS_2 conduction states (Fig. 1). This is consistent with the scanning tunnelling spectroscopy (STS) measurements of MoS_2/WS_2 heterostructures [13].

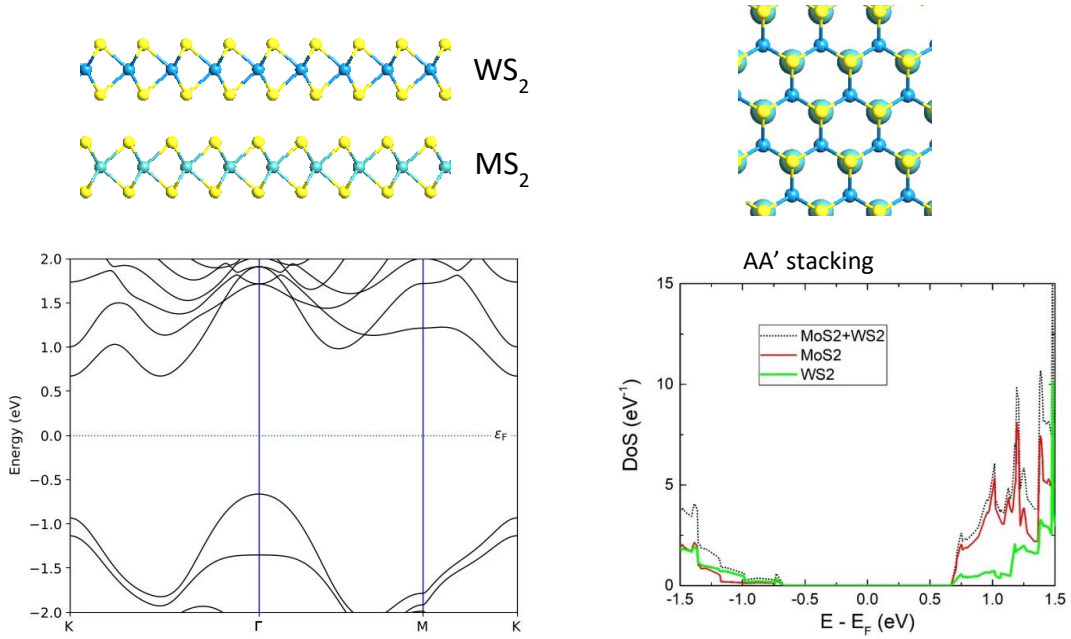


Fig. 1. Top: (left) side-view of MoS_2 and WS_2 monolayers, and (right) top-view of the layers in an AA' stacking configuration. Bottom: (left) band structure, and (right) density of ststes (DoS) of the MoS_2/WS_2 heterstructure.

Density functional theory (DFT) calculations combined with non-equilibrium Green's function (NEGF) formalism allows for investigating the charge transport across the junction and for determining the device properties of vertical interface of MoS_2/WS_2 heterostructure as demonstrated in Fig. 2.

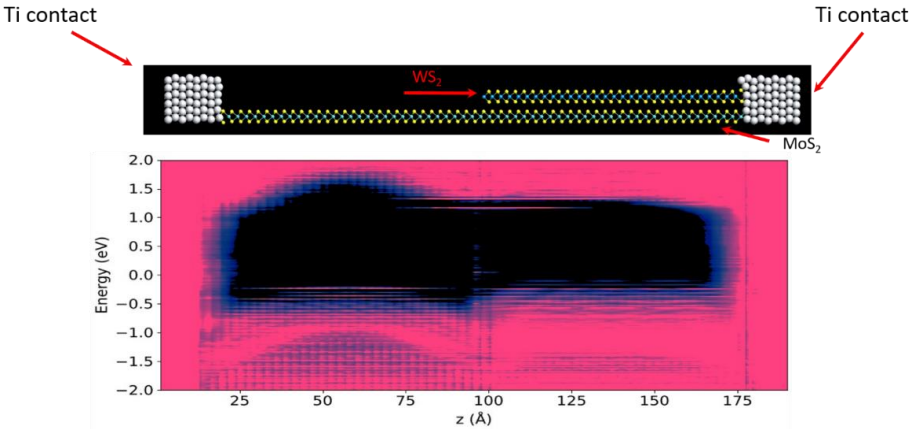


Fig. 2. MoS_2/WS_2 vertical heterstructure (top), and its local DoS at equilibrium (bottom).

Semimetal TMDs provide another posibility of heterojunctions. Platinum diselenide ($PtSe_2$) films demonstrate distinct thickness-dependent electronic structures and physical properties [14, 15]. Although the bulk crystal is a semimetal with an indirect overlap of the conduction and valence bands, monolayer $PtSe_2$ has been revealed to be a semiconductor [16]. The possibility of making semimetal hetero-junctions with uniform chemical bonding at the interface promises the possibility of fabricating ideal Schottky barriers closely mirroring the behaviour of an ideal junction [17, 18]. In this talk, we will discuss how mono-material TMD-based heterostructures could be promising for nanoelectronics applications. Transfer characteristics of two $PtSe_2$-based heterostructure FETs are presented in Fig. 3.

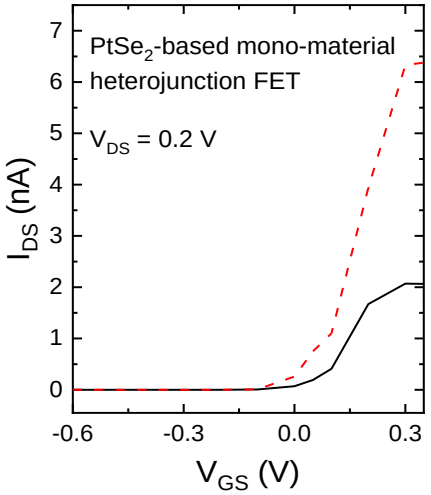


Fig. 3. Transfer characteristic of simulated FET devices based on $PtSe_2$ monomaterial heterstructure.

References

1. Fang, H., et al., Proceedings of the National Academy of Sciences, 2014. **111**(17): p. 6198-6202.
2. He, J., K. Hummer, and C. Franchini, Physical Review B, 2014. **89**(7): p. 075409.
3. Yoon, Y. , et al., Nano Letters, 2011. **11**(9): p. 3768-3773.
4. Choi, M.S., et al., ACS Nano, 2014. **8**(9): p. 9332-9340.
5. Varghese, S.S., et al., Electronics, 2015. **4**(3).
6. Wen, M.Q., et al., Optics Express, 2016. **24**(10): p. 10205-10212.
7. Pu, J., L.-J. Li, and T. Takenobu, Physical Chemistry Chemical Physics, 2014. **16**(29): p. 14996-15006.
8. Marquez, C., et al., 2D Materials, 2020. **7**(2): p. 025040.
9. Kam, K.K. and B.A. Parkinson, The Journal of Physical Chemistry, 1982. **86**(4): p. 463-467.
10. Mak, K.F., et al., Physical Review Letters, 2010. **105**(13): p. 136805.
11. Wang, Q.H., et al., Nature Nanotechnology, 2012. **7**(11): p. 699-712.
12. Tanabe, I., et al., Applied Physics Letters, 2016. **108**(25): p. 252103.
13. Hill, H.M., et al., Nano Letters, 2016. **16**(8): p. 4831-4837.
14. Urban, F., et al., Applied Physics Letters, 2020. **117**(19): p. 193102.
15. Grillo, A., et al., Advanced Functional Materials, 2021. **31**(43): p. 2105722.
16. Ansari, L., et al., npj 2D Materials and Applications, 2019. **3**(1): p. 33.
17. Ansari, L., et al., Nano Letters, 2012. **12**(5): p. 2222-2227.
18. Greer, J.C., et al., Journal of Physics: Condensed Matter, 2018. **30**(41): p. 414003.

Electron-phonon calculations using a Wannier-based supercell approach: applications to the monolayer MoS_2 mobility

Jonathan Backman, Youseung Lee and Mathieu Luisier
Integrated Systems Laboratory, ETH Zürich, Switzerland

Introduction: Recent advances in density functional theory (DFT) have enabled the theoretical study of electron-phonon (el-ph) interactions in various semiconductors. The developed approaches have been extended to the calculation of the carrier mobility of these materials, in particular two-dimensional (2-D) transition metal dichalcogenide (TMD) monolayers. Indeed 2-D TMD-based transistors are widely seen as one of the most promising candidates for future logic switches. Among them, field-effect transistors (FETs) made of MoS_2 monolayers have been shown to reach high on-state currents [1]. To better understand the physics behind carrier transport in MoS_2 FETs, we showcase here an original Wannier-based supercell approach to calculate the MoS_2 mobility and its possible extension towards realistic device simulations. When deriving the method, we found that the inclusion of long-range interactions through an analytic expression of the 2-D Fröhlich [2] has a negligible effect on the total mobility value of MoS_2. We provide an overview of el-ph-limited MoS_2 mobility values by comparing our results to available modeling data [3-11], which shows that our results are in good agreement with the literature [11]. We also put our findings and the available data, in perspective with the experimentally measured values.

Method: As shown in Fig. 1(a), the el-ph coupling elements g(k,q) can be expressed in terms of the electronic Hamiltonian derivatives with respect to atomic displacements Q, i.e. dH/dQ. A DFT finite difference supercell approach can be employed to calculate dH/dQ on a real-space grid, after the DFT Hamiltonians, obtained with VASP [12] for a number of atomic displacements, have been transformed into a localized Wannier basis [13]. By taking into account the symmetry of the system and equivalent atoms, the total number of required displacements can be significantly reduced. Numerical errors are minimized by averaging over symmetrically equivalent dH/dQ elements. We also include a correction, originating from the modified orbital overlaps, to account for the change in basis functions associated with each displacement [14]. Symmetry operations are then used to build a complete set of dH/dQ elements. By construction, the dH/dQ elements are localized and decay as the interacting orbitals are located far away from the displaced atom following the Wannier function properties, as depicted in Fig. 1(b). The Force Constants obtained from the displacement calculations, the Hamiltonian, and the dH/dQ elements can altogether be used to calculate the g(k,q) elements. Subsequently, transport calculations using the linearized Boltzmann Transport Equation (LBTE) or device simulations with the Non-equilibrium Green's function (NEGF) formalism and dedicated scattering self-energies are possible, as illustrated in Fig. 1(c).

Results: The electron and phonon band structures of monolayer MoS_2 are reported in Fig. 2(a). The convergence results for the g(k,q) corresponding to the longitudinal optical phonon are also shown, and compared to an analytical 2-D Fröhlich interaction. Absolute values of g(k,q) calculated on a q-grid for a fixed k-point at the conduction band minimum are displayed in Fig. 2(b) for 4 different phonon modes. The mobility calculations are performed using the iterative LBTE method [6,15] with the g(k,q) calculated on a 201x201x1 k/q-point grid in the full Brillouin Zone, including scattering states up to 0.4 eV above the conduction band minimum. The resulting mobility at different temperatures is plotted as a function of the carrier density in Fig. 2(c). We note that the mobility calculations with and without the 2-D Fröhlich contributions are almost identical (not shown here).

Benchmark Figure 3 summarizes the MoS_2 el-ph-limited mobility values extracted from several *ab initio* studies and compare them to those obtained in this work. In spite of numerous reports about the el-ph carrier mobility of MoS_2, no consensus exists about the "true" value. Mobilities ranging between 130 and 410 $\mathrm{cm}^2/\mathrm{Vs}$ have been published. Recently, Ref. [10] suggested that the origin of those scattered values could be attributed to different pseudopotentials. From our review of the published data, we see no clear trend between the reported mobility values and model parameters such as pseudopotentials, exchange correlation, and perturbation method i.e. density functional perturbation theory (DFPT) or frozen phonons (FP). We here note that other scattering mechanisms, such as surface optical phonon and charged impurity scattering [15], are also known to have a strong impact on the experimentally measured mobility of around 40 $\mathrm{cm}^2/\mathrm{Vs}$ [16,17]. Moreover, it is expected that the MoS_2 intrinsic mobility without screening effects decreases at higher carrier concentrations, due to enhanced intervalley scattering. Based on these observations, the mobility values presented in this work, including carrier and temperature dependencies, are compatible with experimental results.

Outlook: With the calculated quantities constructed on a real space grid, the presented method can be directly used as input to a NEGF-based quantum transport device simulator.

References: [1] K. P. O'Brien et al., *IEDM*, pp.7.1.1 (2021). [2] T. Sohier, et al., *Phys. Rev. B* 94, 085415 (2016). [3] K. Kaasbjerg et al., *Phys Rev B* 85 115317 (2012). [4] X. Li et al., *Phys Rev B* 87 115418 (2013). [5] O .Restrepo et al., *New J. Phys* 16 105009 (2014). [6] W. Li. *Phys Rev B* 92 075405 (2015). [7] T. Gunst et al., *Phys Rev B* 93, 035414 (2016). [8] T. Sohier, et al., *Phys. Rev. Mater.* 2, 114010 (2018). [9] J.-J. Zhou et al., *Comput. Phys. Commun.* 264 107970 (2021). [10] G. Gaddemane et al., *J. Comput. Electron.* 20, 49 (2021). [11] A. Pilotto et al., *Solid-State Electron.* 192, 108295 (2022). [12] G. Kresse et al.r, *Phys Rev B* 54, 11169 (1996). [13] N. Marzari et al., *Phys Rev B* 56, 12847 (1997). [14] T. Frederiksen et al., *Phys Rev B* 75, 205413 (2007). [15] Y. Lee et al., *IEDM*, pp.24.4.1 (2019). [16] B. Radisavljevic et al., *Nat. Mat.* 12, 815 (2013). [17] K. K. H. Smithe et al., *Nano. Lett.* 18, 4516 (2018).

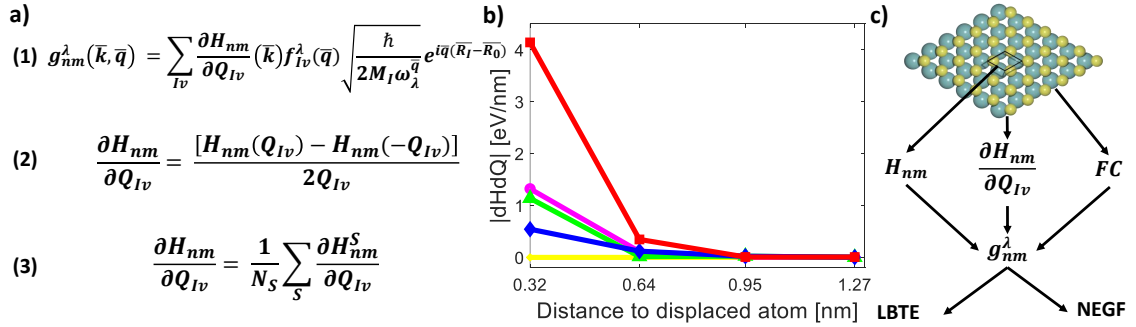


Fig. 1: a) Expressions to calculate the el-ph coupling elements of semiconductors. Here n/m are the orbital indexes, $\lambda(\omega)$ the phonon mode index (energy), $Q_{I\nu}$ represents the displacement of atom I along the direction v, $\overline{k}/\overline{q}$ are the el and ph Bloch vectors, M_I is the displaced ion mass, $\overline{R}$ is a unit cell lattice vector, g^{λ}_{nm} is the el-ph coupling element, f^{λ}_{Iv} the phonon displacement vector, $\frac{\partial H_{nm}}{\partial Q_{Iv}}$ the dH/dQ elements, S indicates dH/dQ elements with symmetry equivalent interactions, and N_S is the number of symmetry operations. b) Example of dH/dQ vs distance of displaced atom data, illustrating the localized nature of the elements. c) Workflow showing the different steps of the method to compute the el-ph coupling elements used in the linearized Boltzmann Transport Equation (LBTE) and in the non-equilibrium Green's function (NEGF) formalism.

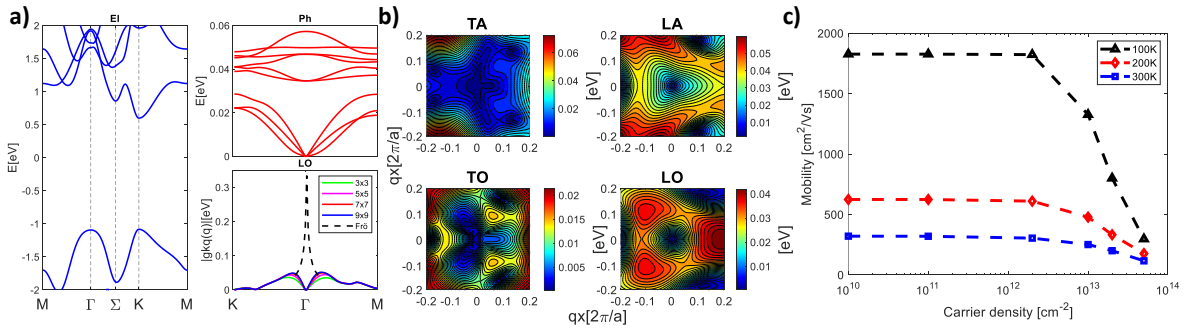


Fig. 2: a) Electronic band structure (left), phonon dispersion (right top), and Fröhlich el-ph coupling elements (right bottom) along the k/q-points of high symmetry. The convergence of the long range interaction for different supercell size is compared to an analytical expression of the 2-D Fröhlich interaction (dashed line). b) Electron-phonon coupling elements calculated on a q-grid for a fixed k-point at the conduction band minimum for 4 different phonon modes: transverse acoustic (TA), longitudinal acoustic (LA), transverse optical (TO), and longitudinal optical (LO). c) 2-D MoS_2 phonon-limited mobility at different temperatures (100, 200, and 300 K) as a function of the carrier density.

Mobility [$cm^2/(Vs)$]				γ	DFT	El-Ph	Ref.
Intrinsic	n = 2e12	1e13	5e13 [cm^{-2}]		Bandstructure	Matrix Elements	
410	400	343	–	1.690	LCAO, LDA	FP	3
130	–	–	–		PW (QE), LDA	DFPT	4
225	–	–	–		PW (QE), PBE	DFPT	5
150	–	–	–		PW (QE), PBE	DFPT + LI	6
380	368	340	244	1.349	LCAO, PBE	FP (9×9)	7
–	–	–	144		PW (QE)	DFPT + LI	8
–	168	–	–	1.378	PW (QE), ONCV-PBE + SOC	DFPT + WI	9
147	–	–	–		PW (QE), ONCV-PBE	DFPT + WI	10
290	275	244	122		PW (QE)	MVDP	11
320	304	251	116	1.582	PW (VASP), PBE	FP (9×9)	This work

Fig. 3: Comparison between the MoS_2 mobility values collected from previous studies and those calculated in this work (marked with blue). The data is classified as a function of the electron density, from low intrinsic density to high (5e13 cm^{-2}). The extracted temperature-dependent exponents γ are provided, when available. The corresponding methods for the calculation of the electronic bandstructure and el-ph matrix elements are also given. Abbreviations: Frozen Phonon (FP), Linear Interpolation (LI), Wannier Interpolation (WI), Multi-Valley Deformation Potential (MVDP), Linear Combination of Atomic Orbitals (LCAO), Plane-Wave (PW), Local density approximation (LDA), Perdew-Burke-Ernzerhof (PBE), Spin-Orbit Coupling (SOC), Optimized Norm-Conserving Vanderbilt (ONCV), Quantum Espresso (QE).

Image-Force Barrier Lowering in Top- and Side-contacted Two-Dimensional Materials

Emeric Deylgat[1,2,3], Edward Chen[4], Massimo V. Fischetti[1], Bart Sorée[3,5,6], and William G. Vandenberghe[1]

[1]Department of Materials Science and Engineering, The University of Texas at Dallas, USA; [2]Department Physics, KU Leuven, Belgium; [3]Imec, Belgium; [4]Corporate Research, Taiwan Semiconductor Manufacturing Company Ltd., Taiwan; [5]Department of Electrical Engineering, KU Leuven, Belgium; [6]Department of Physics, Universiteit Antwerpen, Belgium.

Making low-resistance contacts to 2D materials is challenging and the theory behind contacts is not well-developed, especially the impact of image-force barrier lowering (IFBL) [1,2]. Recently, we calculated the contact resistance in side contacts, accounting for IFBL [3] and showed that using a low-κ dielectric around the 2D material drastically improves contact resistance.

In this work, we analyze IFBL for four different metal-dielectric-2D material configurations as illustrated in Fig. 1 "a"-"d". In "a" we model an edge contact (EC) to the transition metal dichalcogenide (TMD) assuming the dielectric response of the 2D material is the same as that of the dielectric. In "b", we account for the dielectric response of the 2D material assuming MoS_2, in "c" we account for a back-gate, and in "d" we account for a "top" contact to the TMD.

For "a" we use the textbook IFBL expression (Eq. (a)). We previously obtained the IFBL for "b" and "c" using the method of images yielding a Hankel transform [4] (Eq. (b)-(c)). For the top contact we derive the IFBL using the Kontorovich-Lebedev transform which yields Eq (d). The resulting IFBL is plotted in Fig. 2. Figure 3 shows the IFBL as a function of distance x from the interface in the center of the TMD. "a" has the strongest IFBL, "b" has reduced IFBL for very small x, while "c" has much lower IFBL for large x and "d" has even lower IFBL at small x. Figure 4 plots an "effective" dielectric constant which lumps in the contributions of the surrounding dielectrics. The "effective" dielectric constant can be obtained by considering the IFBL as defined in Eq. (a) in all the geometries and calculating from it the dielectric constant as function of x. For configurations "b" and "c", the "effective" dielectric constant approaches ϵ_{2D} for $x < t_{2D}$, whereas for "a" it remains close to the dielectric constant of the environment.

To estimate what the impact of the IFBL is on contact resistance, we consider a conventional Schottky barrier potential $U_S(x) = \frac{eN_D}{2\epsilon}\left(x - x_{dep}\right)^2$ with $x_{dep} = \sqrt{\frac{2\epsilon\phi_s}{eN_D}}$. The tunnel barrier accounting for IFBL for a Schottky barrier height of $\phi_s = 0.3$ eV and $N_D = 10^{10}$ cm^{-2} is illustrated in Fig. 5. We compute the current using the WKB and effective mass approximation ($m^* = 0.5m_e$), and in Fig. 6 we show the contact resistance (Eq. (1)) as a function of doping concentration for contacts "a", "b", "c" and "d" with surrounding SiO_2 ($\epsilon = 3.9\epsilon_0$). We observe that a resistivity of 1 kΩμm is obtained at a doping of 10^{12} cm^{-2} using side contacts. In Fig 8. we show the contact resistance in case of HfO_2 ($\epsilon = 25\epsilon_0$) showing that 1 kΩμm contacts can be achieved at a doping of 10^{14} cm^{-2}. We replot the resistivity data for SiO_2 in Fig. 7 and HfO_2 in Fig. 9 now looking at the ratio between the current with and without. We observe that while using SiO_2 (HfO_2), configuration "a" results in an improvement of up to 50 (2.7), "b" up to 30 (3.6), but the back-gate "c" and the top contact "d" yield smaller improvements of only 15 (3.3) and 10 (1.6), respectively.

In conclusion, we find that the improved IFBL in side-contacts "b" combined with SiO_2 as dielectric can yield contact resistances which are up to 3 times lower compared to top contacts "d". For a HfO_2 dielectric, the contact resistance of side-contact "c" are 2 times lower than top contact "d". However, contacts with a HfO_2 dielectric compared to a SiO_2 dielectric show resistances that are up to 10 times worse. While back-gating results in a higher Schottky barrier compared to contacts without back-gate, at high doping a lower contact resistance can be obtained compared to a top contacted IFBL.

References: [1] M. Mleczko, *et al.*, Nano Letters 19 (9), 6352-6362 (2019).; [2] Allain, A., Kang, J., Banerjee, K. et al. Nature Mater 14, 1195–1205 (2015).; [3] M. Brahma, *et al.*, SISPAD, (2021).; [4] M. Brahma, *et al.*, VLSI-TSA, (2022).

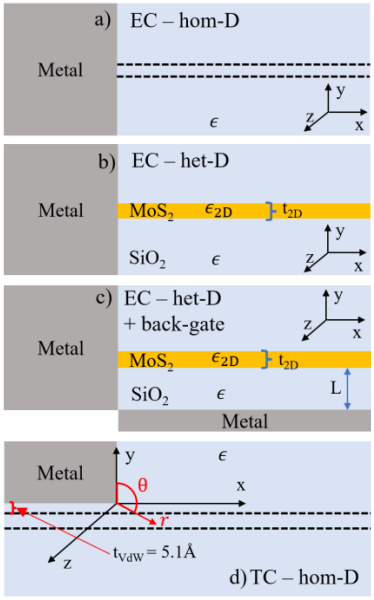


Fig. 1: a) Edge contact (EC) – homogenous dielectric (hom-D); b) EC – heterogeneous D (het-D); c) EC + back gate; d) Top contact (TC) – hom-D.

(a)
$$U_{\text{image}}(x) = -\frac{e^2}{8\pi\epsilon}\frac{1}{2x}$$

(d)
$$U_{\text{image}}(r,\theta) = -\frac{e^2}{8\pi\epsilon}\frac{1}{r}\left[\frac{1}{2} - \frac{2}{3\sqrt{3}} + \int_0^\infty \frac{\cosh(2\alpha\theta)}{\sinh(\alpha\frac{3\pi}{2})}\tanh(\alpha\pi)d\alpha\right]$$

(b) & (c)
$$U_{\text{image}}(x) = -\frac{e^2}{4\pi}\int_0^\infty \hat{V}(Q)J_0(2xQ)QdQ$$

$$\hat{V}(Q) = \frac{\left[\epsilon_{2D}\cosh\left(\frac{1}{2}t_{2D}Q\right) + \epsilon\sinh\left(\frac{1}{2}t_{2D}Q\right)\right]^2 + \left[\epsilon^2\sinh^2\left(\frac{1}{2}t_{2D}Q\right) - \epsilon_{2D}^2\cosh^2\left(\frac{1}{2}t_{2D}Q\right)\right]e^{-2LQ}}{2Q\epsilon_{2D}\epsilon[\epsilon_{2D}\cosh(t_{2D}Q) + \epsilon\sinh(t_{2D}Q)]}$$

(1)
$$\frac{1}{\rho_c} = \frac{2e^2}{h}\int_{-\infty}^{+\infty} dE\left|\frac{df(E)}{dE}\right|\int_{-\infty}^{+\infty}\frac{dk}{2\pi}\exp\left[-2\int_0^{x_{\text{dep}}} dx\sqrt{\frac{2m^*}{\hbar^2}\left(E - \left(U(x) + \frac{\hbar^2k^2}{2m^*}\right)\right)}\right]$$

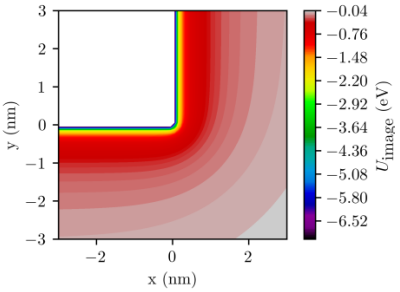


Fig. 2: Image potential energy in the case of d). The surrounding dielectric is SiO_2.

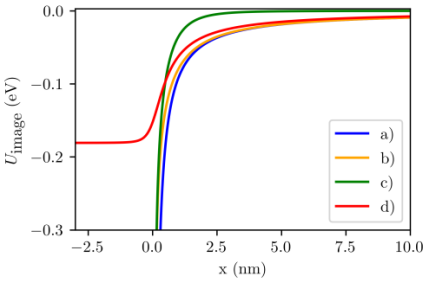


Fig. 3: Image potential energy of the various geometries. The surrounding dielectric is SiO_2.

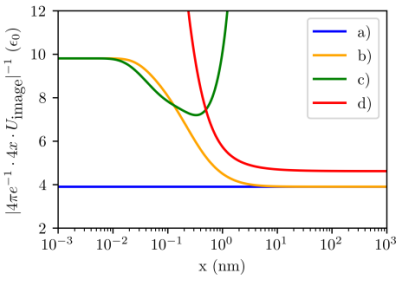


Fig. 4: $\left|4\pi e^{-1}\cdot 4x\cdot U_{image}(x)\right|^{-1}$ vs. x. $\left|4\pi e^{-1}\cdot 4x\cdot U_{image}(x)\right|^{-1}$ is the "effective" dielectric constant. The surrounding dielectric is SiO_2 ($\epsilon = 3.9\epsilon_0$) while $\epsilon_{2D} = 9.8\epsilon_0$.

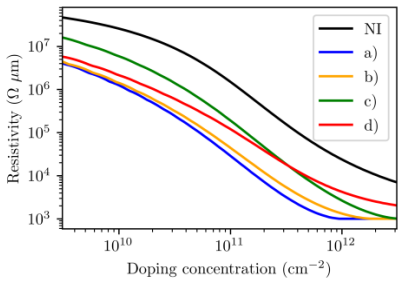


Fig. 6: Contact resistivity vs. doping concentration for a SiO_2 ($\epsilon = 3.9\epsilon_0$) surrounding dielectric: NI - No IFBL, a) - d) with IFBL in various geometries.

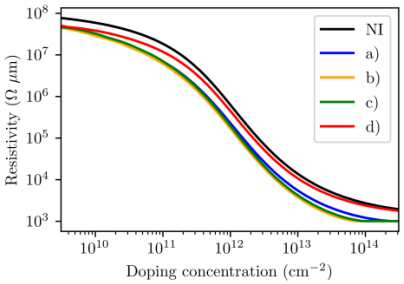


Fig. 8: Contact resistivity vs. doping concentration for a HfO_2 ($\epsilon = 25\epsilon_0$) surrounding dielectric: NI - No IFBL, a) - d) with IFBL in various geometries.

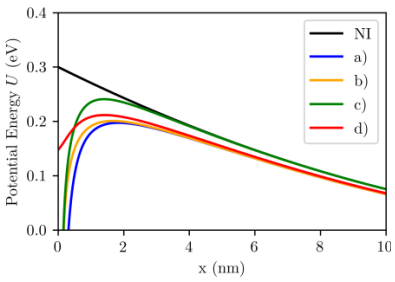


Fig. 5: Potential energy $U(x) = U_S(x) + U_{\text{image}}(x)$: NI - No image force barrier lowering (IFBL), a) - d) with IFBL in various geometries. The surrounding dielectric is SiO_2.

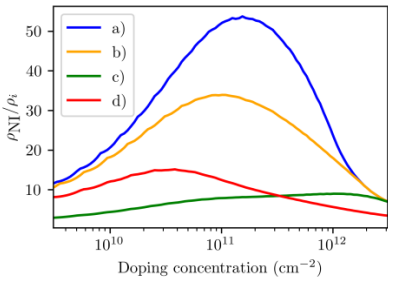


Fig. 7: Ratio of Schottky resistivity w/o IFBL ρ_{NI} and the resistivity of the various geometries ρ_i; i = a), b), c), d) for a SiO_2 ($\epsilon = 3.9\epsilon_0$) surrounding dielectric.

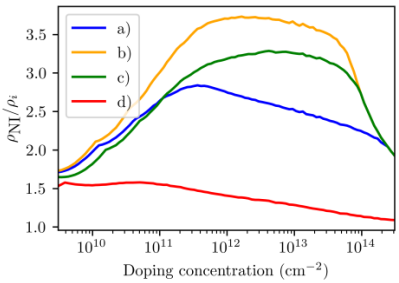


Fig. 9: Ratio of Schottky resistivity w/o IFBL ρ_{NI} and the resistivity of the various geometries ρ_i; i = a), b), c), d) for a HfO_2 ($\epsilon = 25\epsilon_0$) surrounding dielectric.

The impact of electron phonon scattering on transport properties of topological insulators: a first principles quantum transport study

Elaheh Akhoundi
imec
KU Leuven, Dept. of Physics and Astronomy
Leuven, Belgium
elaheh.akhoundi@imec.be

Michel Houssa
imec
KU Leuven, Dept. of Physics and Astronomy
Leuven, Belgium
michel.houssa@imec.be

Aryan Afzalian
imec
Leuven, Belgium
aryan.afzalian@imec.be

***Abstract*—Using first-principles calculations and non-equilibrium Green's function, we study the topologically protected carrier transport in the edges of topological insulator ribbons. We investigate the effects of electron-phonon interactions on the edge state transport. We observed that in topological insulator with small bulk gaps, electron phonon scattering results in the bulk states broadening into the bulk gap. This leads to the destruction of the dissipationless transport. However, if the transport is restricted to the protected states, a higher immunity to electron-phonon scattering can be achieved. This can lead to the design of topological insulator field effect transistors operating based on scattering modulation, benefiting from a channel material with strong electron-phonon coupling.**

Keywords—semiconductor device modelling, topological insulators, 2D materials, NEGF

I. Introduction

Topological insulators (TIs) have attracted considerable attention due to their unique properties. 2D topological insulators possess spin-polarized gapless edge states that are robust against elastic backscattering [1-2]. The absence of backscattering is attributed to the bulk electronic structure and time-reversal symmetry. Distinguished by a non-zero Z2 topological invariant, the edge states are protected even when large level of imperfections such as vacancy defects, phonons etc are present [3-4].

The observation of the quantized conductance of the protected states is an essential step in the study of the possible applications of TIs. This has proven to be challenging in the lab and the experimental success have been limited to very low temperature or very short channels [5-6]. This has been attributed to various backscattering mechanisms, such as electron-electron interactions, defects, phonon scattering [7]. Here, we study the effects of electron-phonon coupling on the topologically protected edge conduction using a first principles description and non equilibrium Green's function (NEGF) formalism implemented in our ATOmistic MOdeling Solver (ATOMOS) [8,9]. Using ATOMOS, we are able to conduct transport simulations on wide TI ribbons, which contain heavy atoms and include spin-orbit coupling effects. To decrease the computational costs, we used a mode-space transformation to reduce the size of the Hamiltonian and overlap matrices [10,11]. In Section II, we summarize the computational methods used in our simulations. We discuss the results in Section III and draw conclusions in Section IV.

II. Method

ATOMOS operates in a NEGF framework. Density functional theory (DFT) simulations are carried out to compute the electronic properties of a material, including Hamiltonian matrix elements. Here we focus on two widely studied topological insulators, stanene and bismuthene. The former has a small bulk gap of 0.17eV [12] while the latter possess a larger bulk gap of 0.5 eV [4]. By choosing these two materials, we can compare the impact of electron phonon interactions on protected and non-protected states.

The OpenMX package [13-14] is used to perform DFT simulations and to extract the matrix elements. The generalized gradient approximation (GGA) implementation of DFT with the Perdew-Burke-Ernzerhof (PBE) exchange-correlation (XC) functional is employed. The Brillouin zone was sampled by a 16x1x1 Monkhorst-Pack grid. All the structures are fully relaxed with atomic force convergence criterion of 10^{-3} eV/Å. A vacuum of 15Å is introduced in the non-periodic directions. The spin-orbit coupling effects are included.

The Hamiltonian and overlap matrix elements are imported to ATOMOS. The non-orthogonal Hamiltonian matrix for the full device is then constructed. Dissipative transport calculations are carried out using NEGF and the self-consistent Born approximation [15,16]. The complete simulation procedure is detailed in [9,17].

III. Results

Fig. 1(a) shows the band structure of a 4 nm wide zigzag stanene ribbon. The spin-resolved contribution of the edge atoms to the band structure is shown as well. The topologically-protected edge states possess a nearly linear dispersion as well as a helical spin texture. The spin texture ensures that intra-edge backscattering is forbidden. For the states in the bulk gap, there is a noticeable edge localization of the wavefunction. However, the wavefunction associated with a state outside the bulk gap is not localized at the edges.

Fig. 1(b) compares the band structure calculated by a real-space Hamiltonian (blue dots) to the band structure computed using a reduced-size, mode-space Hamiltonian (red dots). The mode-space method is used to construct the band structure in an energy window of interest. Here the energy window includes edge states as well as a few bulk bands. As the Fermi level lies inside the bulk gap, carrier transport takes place through the dissipationless edge bands. By shifting the

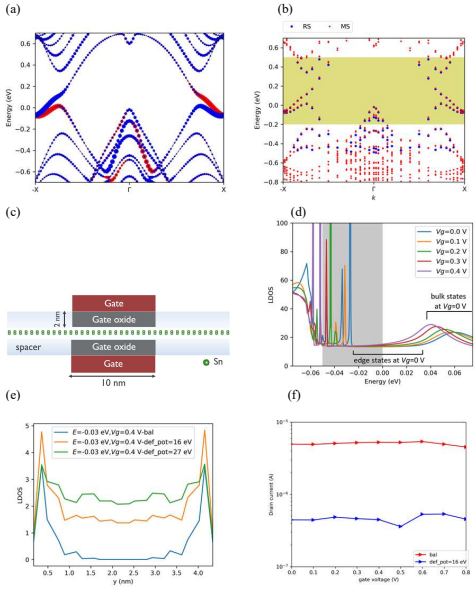


Fig. 1. (a) the spin-resolved band structure projected on the edge atoms obtained for a 4 nm wide zigzag stanene nanoribbon with Fluorine edge passivation. Blue dots show spin-up states and red dots represent the spin-down states. (b) the band strcuture calculated by the original (RS) and mode-space (MS) Hamiltonians. The shaded area shows the energy window of interest. (c) the sketch of the device structure under study (d) the calculated LDoS for different gate voltages. The shaded area exhibits the Fermi window. (e) LDoS across the width of the ribbon at E= -0.03 eV and Vg=0.4 V. (f) IV curves for the ballistic limit as well as the dissipative case with a deformation potential of 16 eV.

Fermi level to the bulk bands, bulk states also contribute to the current. The shift in the Fermi level from edge states to the bulk states can result in scattering modulation, which is one of the switching mechanisms proposed for a TI field effect transistor [12].

The Fermi level can be modulated using a gate, as shown in Fig. 1(c). The LDoS (local density of states) taken from the middle of the channel of a TI subjected to a gate voltage in the ballistic limit is shown in Fig. 1(d). The source-drain bias is 0.05 V. The shaded area in Fig. 1(d) represents the Fermi window. As the gate voltage increases, the bands move down and at V_G =0.4 V, the entire Fermi window is covered by edge states. for higher gate voltages, the bulk states in the conduction band will move down to inside the Fermi window.

The LDoS across the width of ribbon at energy of -0.03 eV for V_G=0.4 V is denoted in Fig. 1(e). In the absence of electron phonon interactions, the state is fully localized at the edges. However, as the electron phonon scattering is taken into account, the broadening of the bulk states into the bulk gap results in considerably higher contribution of the bulk atoms to the LDoS of the considered Fermi window. It is clear from Fig. 1(e) that even a small deformation potential of 16 eV for acoustic phonon scattering destroys the edge localization and leads to strong degradation of the current as shown in Fig. 1(f). As the degradation is not merely a product of inter-edge scattering, increasing the width of the ribbon will not significantly improve the transport properties.

Figures 2(a) and 2(b) illustrate the energy band structure of a 7 nm wide bismuthene zigzag nanoribbon,

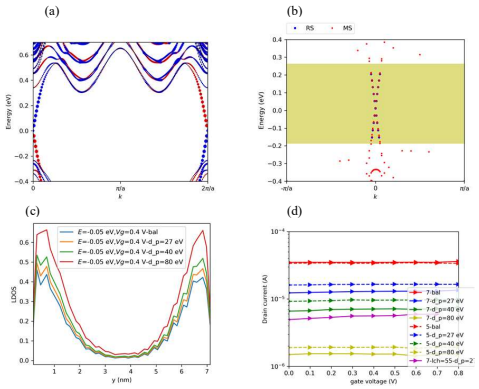


Fig. 2. (a) the spin-resolved energy band structure pojected on the edge atoms of 7 nm wide zigzag bismuthene nanoribbon with H edge passivation.(b) the band strcuture in a target energy window calculated by RS and MS Hamiltonians. (c) LDOS accross the ribbon at E=-0.05 eV for the ribbon under gate voltage. Different strength of electon-phonon coupling are compared by considering different deformation potentials as well as the ballistic limit (bal). (d) IV curves for ZZ bismuthene ribbons of widths 5nm (5) and 7nm (7) for several deformation potential values. The length of all the strcutures is 24 nm except for one structure with Lch=55nm (Ltot=70nm).

calculated by the real-space and mode-space Hamiltonians, respectively. The energy window of interest only contains the edge bands to gain an insight into the effect of phonon scattering on the topological protection. Furthermore, by tunning the Fermi level and the drain bias the bulk states will not participate in the transport, due to the large bulk gap. A source-drain bias of 0.2 V is considered here. Fig. 2(c) demonstrates that even a deformation potential as large as 40 eV will not destroy the edge localization.

The IV curves for the TI under gate voltage for different deformation potentials are presented in Fig. 3(d). The total length of the structure is 24 nm, unless mentioned otherwise. As the deformation potential increases to 40 eV, the drop in current is not as significant as that calculated for the stanene case. However, if the deformation potential is set to 80 eV, the current declines by an order of magnitude. The IV curves for a bismuthene nanoribbon with a width of 5nm are also plotted in Fig. 2(d). It is apparent that even for a smaller ribbon width, there is similar immunity to backscattering. Additionally, we increased the channel length to 55nm, resulting in a full device length of 70nm. The IV curve for deformation potential of 27 eV is shown in Fig. 2(d) (purple line). The drop in current is smaller than one order of magnitude, supporting the idea of prohibited backscattering.

IV. CONCLUSIONS

We observed that electron-phonon interactions can result in the degradation of the protected transport in a topological insulator. In stanene with a small bulk gap of 0.17 eV, electron phonon coupling results in the extension of the bulk states into the bulk gap. Therefore, the edge localization of the states within the bulk gap is destroyed by a deformation potential as small as 16 eV. On the other hand, if we only consider the transport through the protected states of bismuthene (bulk gap~0.5 eV), a higher immunity to the electron phonon scattering is obtained. The current degradation due to a deformation potential of 27 eV at a higher drain-source bias is improved significantly compared to that of stanene.

REFERENCES

[1] Hasan, M. Zahid, and Charles L. Kane. "Colloquium: topological insulators." *Reviews of modern physics* , vol. 82, no. 4, p. 3045, 2010.

[2] Bansil, Arun, Hsin Lin, and Tanmoy Das. "Colloquium: Topological band theory." *Reviews of Modern Physics*, vol. 88, no. 2, p. 021004, 2016.

[3] Tiwari, Sabyasachi, Maarten L. Van de Put, Bart Sorée, and William G. Vandenberghe. "Carrier transport in two-dimensional topological insulator nanoribbons in the presence of vacancy defects." *2D Materials* vol. 6, no. 2, p. 025011, 2019.

[4] Pezo, Armando, Bruno Focassio, Gabriel R. Schleder, Marcio Costa, Caio Lewenkopf, and Adalberto Fazzio. "Disorder effects of vacancies on the electronic transport properties of realistic topological insulator nanoribbons: The case of bismuthene." *Physical Review Materials,* vol. 5, no. 1, p. 014204, 2021.

[5] Konig, Markus, Steffen Wiedmann, Christoph Brune, Andreas Roth, Hartmut Buhmann, et al. "Quantum spin Hall insulator state in HgTe quantum wells." *Science,* vol. 318, no. 5851, pp. 766-770, 2007.

[6] Wu, Sanfeng, Valla Fatemi, Quinn D. Gibson, Kenji Watanabe, Takashi Taniguchi, et al. "Observation of the quantum spin Hall effect up to 100 kelvin in a monolayer crystal." *Science,* vol. 359, no. 6371, pp. 76-79, 2018.

[7] Vannucci, Luca, Thomas Olsen, and Kristian S. Thygesen. "Conductance of quantum spin Hall edge states from first principles: The critical role of magnetic impurities and inter-edge scattering." *Physical Review B,* vol. 101, no. 15, p. 155404, 2020.

[8] A. Afzalian and G. Pourtois, "Atomos: An atomistic modelling solver for dissipative dft transport in ultra-scaled hfs2 and black phosphorus mosfets," in 2019 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), pp. 1–4, 2019.

[9] A. Afzalian, "Ab initio perspective of ultra-scaled CMOS from 2d material fundamentals to dynamically doped transistors," npj 2D Materials and Applications, vol. 5, no. 1, 2021.

[10] A. Afzalian , J. Huang, H. Ilatikhameneh, J. Charles, D. Lemus, J. Bermeo Lopez, et al. "Mode space tight binding model for ultra-fast simulations of III-V nanowire MOSFETs and heterojunction TFETs." In *2015 International Workshop on Computational Electronics (IWCE)*, pp. 1-3. IEEE, 2015.

[11] A. Afzalian, T. Vasen, P.Ramvall, T.-M. Shen, J. Wu and M. Passlack, "Physics and performance of III-V nanowire broken-gap heterojunction TFETs using an efficient tight-binding mode-space NEGF model enabling million-atom nanowire simulations", J. Phys. Condens. Matter, vol. 30, no. 25, 254002 (16pp), 2018. https://doi.org/10.1088/1361-648X/aac156.

[12] W. Vandenberghe, and M. Fischetti. "Imperfect two-dimensional topological insulator field-effect transistors." *Nature communications* vol. 8, no. 1, pp. 1-8, 2017.

[13] Ozaki, Taisuke. "Variationally optimized atomic orbitals for large-scale electronic structures." *Physical Review B,* vol. 67, no. 15, p. 155108, 2003.

[14] Kotaka, Hiroki, Fumiyuki Ishii, and Mineo Saito. "Rashba effect on the structure of the Bi one-bilayer film: Fully relativistic first-principles calculation." *Japanese Journal of Applied Physics* 52.3R, p. 035204, 2013.

[15] A. Afzalian. "Computationally efficient self-consistent born approximation treatments of phonon scattering for coupled-mode space non-equilibrium Green's function." *Journal of Applied Physics,* vol. 110, no. 9, p. 094517, 2011.

[16] A. Afzalian, G. Doornbos, T.-M. Shen, M. Passlack and J. Wu, "A High-Performance InAs/GaSb Core-Shell Nanowire Line-Tunneling TFET: An Atomistic Mode-Space NEGF Study", IEEE J. of Electron Dev. Society, Nov. 2018, DOI: 10.1109/JEDS.2018.2881335.

[17] A. Afzalian, E. Akhoundi G. Gaddemane, R. Duflou and M. Houssa, "Advanced DFT–NEGF Transport Techniques for Novel 2-D Material and Device Exploration Including HfS2/WSe2 van der Waals Heterojunction TFET and WTe2/WS2 Metal/Semiconductor Contact," in IEEE Transactions on Electron Devices, vol. 68, no. 11, pp. 5372-5379, Nov. 2021, doi: 10.1109/TED.2021.3078412

Theoretical Study of Carrier Transport in Two-dimensional Transition Metal Dichalcogenides for Field-Effect Transistor Applications

Sanjay Gopalan, Maarten L. Van de Put, and Massimo V. Fischetti
Department of Materials Science and Engineering, The University of Texas at Dallas, Texas, USA

Theoretical methods are critical for evaluating two-dimensional materials as possible channel materials in future field-effect transistors (FETs). Most of the modelling attempts concentrate on electronic transport in ideal free-standing layers, ignoring the dielectric environment's effect on transport characteristics. We study the effect of the dielectric environment on carrier transport in two-dimensional (2D) transition metal dichalcogenides (TMDs) by extending our Monte-Carlo model for a free-standing monolayer to include dielectric screening and remote-phonon scattering in a double-gate.

There are very few studies in the literature that look at how the dielectric environment affects the transport properties of 2D TMDs [1-3]. However, these studies have only considered a single TMD (namely, MoS_2) or have been based on simplifications of the problem obtained, for example, by ignoring the full hybridization of all phonon-like and plasmon-like modes, of free-carrier screening effects, and using approximate dispersions of the hybrid modes. In our work, we study the scattering of electrons with the coupled (hybrid) optical-phonon/plasmon excitations that are present at the interfaces in a field-effect transistors with a channel consisting of a monolayer of a polar semiconductor (TMDs) with top and bottom gates. We consider the full hybridization of all modes and free-carrier screening.

Of particular interest is the work done by Hauber and Fahy [3]. They present an outstanding analysis of the plasmon/optical-phonon excitations in bulk III-V compound semiconductors and supported and gated thin MoS_2 layers. However, they fail to take into account of different gate insulators and the polar semiconductors. We preform an extensive study of the electron mobility in supported and/or gated monolayer TMDs in the presence of "remote-phonon" scattering, extending the study of the room-temperature electron mobility not only in MoS_2 but also in WS_2, $MoTe_2$, and of the 300 K hole mobility in WSe_2, with all TMDs assumed to be supported by SiO_2 and with SiO_2, HfO_2, Al_2O_3, AlN, and hBN as gate insulators with an equivalent oxide thickness of 0.7 nm.

Among the various supported and gated TMDs considered, the best electron mobility, 782 $cm^2/(V.s)$, is obtained for the SiO_2/WS_2/hBN system. An excellent hole mobility, 387 $cm^2/(V.s)$ is also predicted for the SiO_2/WSe_2/hBN. These extended results confirm the general trend observed previously for MoS_2[4]: The carrier mobility decreases almost monotonically with increasing dielectric constant of the gate insulator, with two exceptions: **1.** The beneficial effects of dielectric screening of the 'out-of-plane' field lines are seen for hBN, thanks to its relatively low ionic polarization and the high phonon frequencies resulting from the light weight of the B and N ions. **2.** On the contrary, resonance effects among the optical phonons of the substrate, of the TMD layer, and of the top oxide result in a low carrier mobility when AlN and/or Al_2O_3 are taken as gate insulators.

Figure 1 shows the mobility results for various combinations. Looking at the Fig. 1(a), one notices that the calculated 300 K electron mobility in SiO_2-supported monolayer MoS_2 decreases as the gate insulator is changed from hBN to HfO_2, but it reaches a particularly high value for hBN and a very low value for Al_2O_3. A similar behavior is seen also for the 300 K electron mobility of $MoTe_2$ (Fig. 1(b)), for the hole mobility in WSe_2 (Fig. 1(d)) and, for electrons in WS_2 (Fig. 1(c)), when AlN is taken as gate insulator. In all cases, hBN exhibits the best mobility, even higher than in free-standing monolayers, thanks to the dielectric screening of the out-of-layer field lines and the weak remote-phonon scattering due to its weak polar nature and light ions (resulting in high phonon frequencies, decoupled from all other excitations).

We then extend our model to simulate a 2D material-based field-effect transistor, considering various two-dimensional TMDs as the channel material. [4][5]

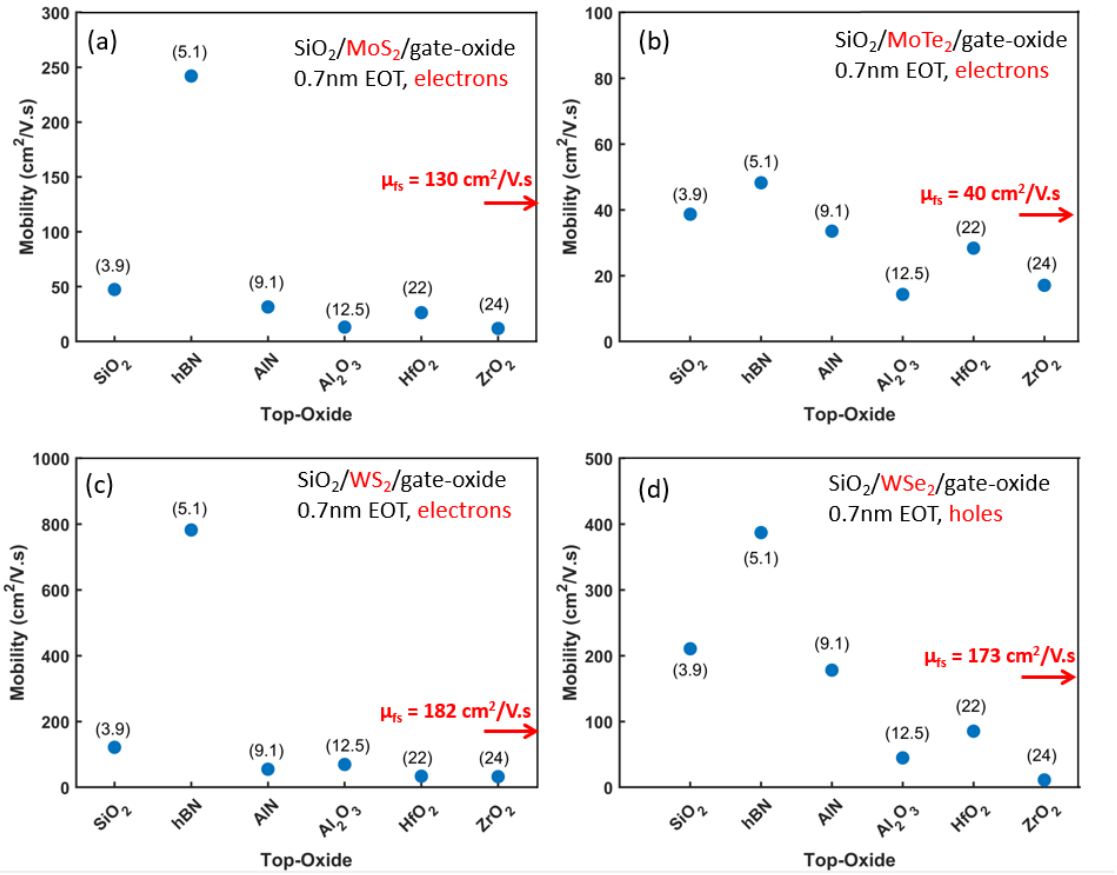


Figure 1. **(a)**: Calculated 300 K electron mobility in the TMD channel of a SiO_2/MoS_2/gate-oxide structure with different gate oxides (ordered with increasing static dielectric constant – shown in parentheses -- from left to right). Note that the mobility generally decreases as the static dielectric constant of the gate oxide increases, but phonon resonances cause 'dips' in the cases of Al_2O_3 and, to a smaller extent, ZrO_2. Moreover, the absence of low-energy phonons in hBN results in a mobility larger than in free-standing MoS_2 (shown by the red arrow), thanks to the beneficial effect of dielectric screening of the bulk phonon. **(b)**: The same, but for the $SiO_2/MoTe_2$/gate-oxide structure and **(c)**, for the SiO_2/WS_2/gate-oxide structure. One can notice a similar behavior. Finally, frame **(d)** shows the calculated 300 K hole mobility in the SiO_2/WSe_2/gate-oxide system.

References

[1] Lang Zeng, Zheng Xin, Shaowen Chen, Gang Du, Jingfen Kang, and Xiaoyan Liu, *Remote phonon and impurity screening effect of substrate and gate dielectric on electron dynamics in single layer MoS_2*, Appl. Phys. Lett. **103**, 13505 (2013).

[2] Nan Ma and Debdeep Jena, *Charge Scattering and Mobility in Atomically Thin Semiconductors*, Phys. Rev. X **4**, 011043 (2014).

[3] Anna Hauber and Stephen Fahy, *Scattering of carriers by coupled plasmon-phonon modes in bulk polar semiconductors and polar semiconductor heterostructures*, Phys. Rev. B **95**, 045210 (2017).

[4] Maarten L Van de Put, Gautam Gaddemane, Sanjay Gopalan, and Massimo V Fischetti. *Effects of the dielectric environment on electronic transport in monolayer MoS_2: Screening and remote phonon scattering*. In 2020 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), pages 281–284. IEEE.

[5] Gautam Gaddemane, Maarten L Van de Put, William G Vandenberghe, Edward Chen, and Massimo V Fischetti, *Monte Carlo analysis of phosphorene nanotrasistors*, arXiv: 2007.14940, 2020.

Massively Parallel FDTD Full-Band Monte Carlo Simulations of Electromagnetic THz pulses in p-doped Silicon at Cryogenic Temperatures

C. Jungemann,[a] F. Meng,[b] M. D. Thomson,[b] and H. G. Roskos[b]
[a] Chair of Electromagnetic Theory, RWTH Aachen University, 52056 Aachen, Germany
[b] Physikalisches Institut, Goethe-Universität Frankfurt, 60438 Frankfurt am Main, Germany

Intense electromagnetic THz pulses are used to probe the properties of semiconductor materials. The response of the particle gas to the strong electric field is nonlinear and results in the generation of higher harmonics, which are due to the nonparabolic band structure and energy-dependent scattering [1]. Here we theoretically investigate experiments where an intense THz pulse passes through a weakly p-doped silicon layer (thickness 275 µm) where the beam is normal to the [001] surface of the layer and the linearly polarized electric field is aligned with the [100] direction. The sample is kept at a temperature of 10 K and the hole density is much lower than the acceptor concentration of $5 \times 10^{16}\,\mathrm{cm}^{-3}$ due to impurity freeze-out. While the pump field dependence indicates that additional carriers are generated during the pulse (attributed to impact ionization [2]), here we assume a constant total population of band carriers, which still captures the main aspects of the harmonic generation. In measurements the intensity of the 3rd harmonic is many orders of magnitude lower than the fundamental one confirming that the hole density is indeed very low.

The low particle density allows one to decouple the FDTD pump-pulse propagation and Monte Carlo response of the hole gas, which greatly reduces the computational cost of the simulations. First, the electromagnetic pulse is simulated for a vanishing hole density by a 1D FDTD solver [3, 4] and the electric field in the silicon layer is recorded as a function of location and time. A 1D approach is sufficient because of the rather large diameter of the beam and that the measured harmonics are dominated by the high intensity at the beam center. Second, the hole velocity is simulated for this electric field by a full-band Monte Carlo code [5], where the electric field due to the very low space charge density can be neglected [4], and the particles can be simulated independently. 275 jobs are started in parallel for the boxes of the real space grid in the silicon region. Third, a FDTD simulation with the current density obtained in the second step is performed, where the electric field at the fundamental frequency caused by the current density is much smaller than the one due to the THz source. The transmitted electric field is Fourier transformed yielding the spectrum of the higher harmonics. If self-consistency is required, steps 2 and 3 can be repeated with the updated values until convergence is obtained.

A fine grid of 1 µm and a time step of 1.67 fs is used to resolve the higher harmonics. The electric field of the source pulse is given by a sinusoidal carrier wave with a frequency of 1.29 THz and a Gaussian envelope function (Fig. 1). A fraction of the pulse is reflected at both interfaces between the sample and vacuum leading to a partially standing wave in the vacuum at the source side and in the sample. The maximum value of the magnitude of the electric field over time is shown in Fig. 2 for a source amplitude of $28\,\mathrm{kV\,cm}^{-1}$, and the maximal electric field strongly varies in the sample. The intensity of the transmitted field is the absolute square of the Fourier transform and at odd multiples of the fundamental frequency additional peaks occur, when the response of the hole gas is included in the simulation (Fig. 3). The generation of the higher harmonics is due to the nonparabolic band structure, which is calculated by the nonlocal empirical pseudopotential method [6] and discretized by an unstructured tetrahedral mesh with a high resolution near the Γ-point (Fig. 4), and scattering [5]. The total CPU time is about 3 months and the wall clock time about 8 hours for the simulation of $1.375 \cdot 10^8$ particles. The resultant noise is very low and even the 9th harmonic is clearly visible in Fig. 3.

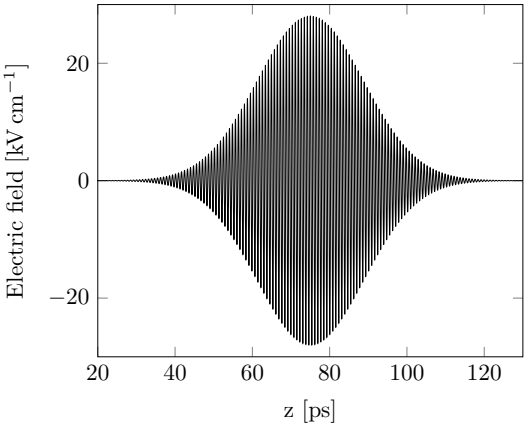


Figure 1: Electric field of the THz pulse generated by the source for an amplitude of $28\,\mathrm{kV\,cm}^{-1}$.

In Fig. 5 the result of a transient bulk Monte Carlo simulation is shown (i.e. calculated at a single spatial point), where the electric field is given by the source field. The intensity is calculated based on the time-dependent velocity. The resultant ratio of the 3rd and 5th harmonic is much smaller than in the case of the FDTD full-band Monte Carlo simulation,

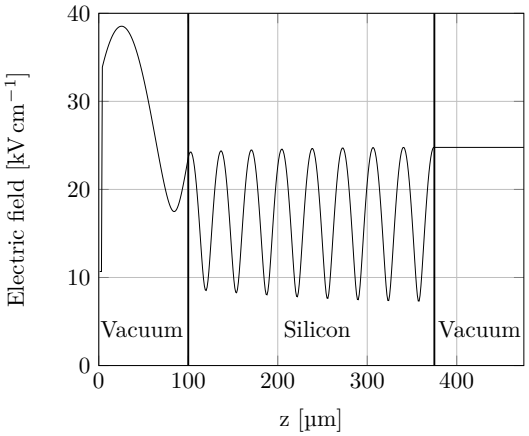


Figure 2: Maximum of the pump-pulse electric field over time versus location for a source amplitude of 28 kV cm^{-1}.

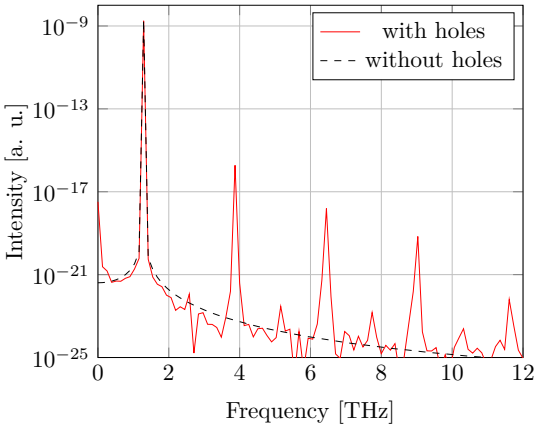


Figure 3: Intensity of the transmitted electric field for a source amplitude of 28 kV cm^{-1} and frequency of 1.29 THz.

showing that a simple bulk Monte Carlo simulation is not sufficient to quantitatively describe the generation of higher harmonics in the silicon layer and the full wave propagation has to be taken into account.

[1] F. Meng et al., "High-harmonic generation from weakly p-doped Si pumped with intense THz pulses," in *2021 46th International Conference on Infrared, Millimeter and Terahertz Waves (IRMMW-THz)*, 2021, pp. 1–1.

[2] F. Meng et al., "Intracavity third-harmonic generation in Si:B pumped by intense Terahertz pulses," *Phys. Rev. B*, vol. 102, p. 075205, Aug 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.102.075205

[3] K. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Transactions on Antennas and Propagation*, vol. 14, no. 3, pp. 302–307, 1966.

[4] K. J. Willis et al., "Global modeling of carrier-field dynamics in semiconductors using EMC–FDTD," *Journal of Computational Electronics*, vol. 8, no. 2, p. 153, Aug 2009.

[5] C. Jungemann and B. Meinerzhagen, *Hierarchical Device Simulation: The Monte-Carlo Perspective*, ser. Computational Microelectronics. Wien, New York: Springer, 2003.

[6] E. Ungersboeck et al., "The effect of general strain on the band structure and electron mobility of silicon," *IEEE Trans. Electron Devices*, vol. 54, no. 9, pp. 2183–2190, 2007.

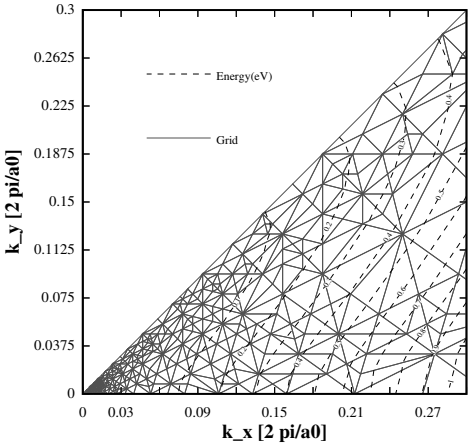


Figure 4: K-space grid of the heavy-hole band for $k_z = 0$ [5].

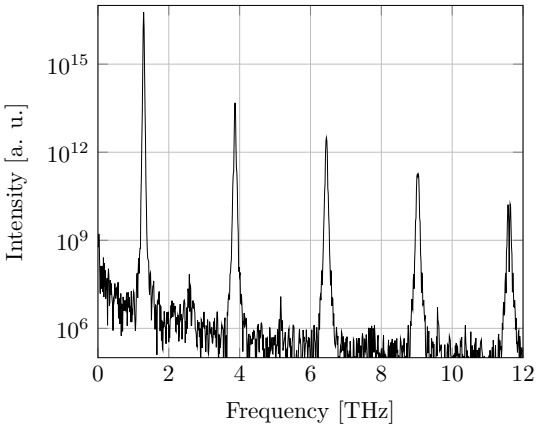


Figure 5: Intensity of the bulk velocity for an amplitude of 24 kV cm^{-1} and frequency of 1.29 THz.

TCAD simulation of microwave circuits: the Doherty amplifier

S. Donati Guerrieri[1], E. Catoggio[1], F. Bonani[1]
[1]Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, Italy

Abstract

Power amplifiers (PAs) for next generation of communication systems are expected to operate at higher frequency and bandwidth to support the growing data rates. The Doherty Amplifier (DA) is one of the most promising circuits for the development of high efficiency PAs: its inherent structure, exploiting two interacting active devices, requires nonlinear mixed-mode TCAD analysis with multiple devices. In this paper we present for the first time the TCAD simulation of a DA. The TCAD analysis has been carried out exploiting our in-house Harmonic-Balance based drift-diffusion simulator allowing for large signal mixed-mode analysis. We demonstrate for the first time that TCAD simulations are mature to assist the design of complex stages requiring multidevice large signal analysis.

Index Terms

Nonlinear TCAD, Harmonic Balance, High Efficiency Power Amplifiers, Doherty Power Amplifier

I. Introduction

One of the main challenges in the development of 5G/6G telecom front-ends is the design of broadband PAs featuring high efficiency from back-off to saturation, operating with high Peak-to-Average Power Ratio (PAPR) signals. Various approaches have been proposed for high efficiency PAs, most of them based on circuits with more that one (typically two) interacting active devices[1], [2]. In the DA[3], [4], two active devices (FETs) from the same technology are combined so that one (the so-called peak or auxilliary - AUX) acts like an active load for the other (the main - MAIN). Ideally, a well designed DA should exhibit a nearly constant efficiency at least in a range of 3dB output power back-off (OBO) condition [3], even though practical realizations show poorer performance due to the difficult design of the output coupling and matching network (OCMN)[5]. In fact, the AUX and MAIN interact through the OCMN, differently from the usual parallel stage PA, where devices are isolated. In the DA, the AUX and MAIN operating conditions depend on the input power drive, which calls for a large signal analysis of the two devices concurrently with the OCMN. Circuit level analysis is difficult since accurate device models from deep class C to class AB are often not available from foundry design kits. TCAD simulations are the ideal way to analyze and optimize the DA, especially regarding two critical points: the AUX output capacitance model in large signal operation, and the AUX behavior close to the threshold (turn-on).

II. TCAD simulation of the Doherty stage

To test the TCAD mutidevice Harmonic Balance mixed-mode simulator, we have designed a preliminary simplified DA at 12 GHz exploiting two MESFET GaAs epitaxial devices with 0.5 μm gate length, 2×10^{17} cm^{-3} channel doping, 1.5 μm source/drain separation [6], [7], and 1 mm gate periphery. Fig. 1 shows the simulated DA with the two MESFETs discretized for TCAD analysis (≈ 6500 grid nodes and 10 harmonics) and the embedding OCMN. The AUX operates as a class C stage: at lower input power it is off and acts as an open load; at higher power it turns on, operating as a high efficiency PA while providing the correct load to maintain the MAIN at maximum efficiency. The AUX gate bias choice is critical for the correct AUX turn on: we have optimized it at $V_{GS,AUX} = -5.15$ V. The MAIN is biased in deep class AB ($V_{GS,MAIN} = -3$ V). The drain bias is 8 V. The OCMN is made of a combination of transmission lines; the MAIN equivalent load is $2\,\mathrm{Re}(Z_{opt})$ in back-off (AUX off) and $\mathrm{Re}(Z_{opt})$ when both the devices are on (Z_{opt} is the optimum load impedance). The output MAIN and AUX capacitances have been tuned out by parallel equivalent negative ones, whose value is extracted from the small signal Y matrix. At higher input power, output MESFET capacitances change and a significant de-tuning can appear. Two swept voltage generators with internal 50 Ω impedance provide the input power: the AUX available input power is set $\sqrt{2}$ more than the MAIN. Input ports are unmatched in this preliminary analysis.

Fig. 2 shows an example of the obtained results. The DC component of the device electron concentration is shown at three increasing powers, from back-off to saturation. In back-off the AUX (to the right) has a depleted channel, which turns-on at increasing input power due class C self-biasing. At saturation the two devices exhibit nearly the same channel populations. Fig. 3 shows the AUX and MAIN dynamic load lines. In back-off the AUX is off. At intermediate power AUX turns on, while MAIN comes close to its maximum efficiency (the DLL of the MAIN reaches the maximum swing allowed by the knee voltage). Finally, in saturation the two devices act very similarly, both contributing to the maximum output power. Fig. 4 shows the DA RF performance: on the left, the output power is shown along with the individual contributions of the AUX and MAIN. Here we also notice the AUX turn on, anticipated with respect to the ideal 3 dB OBO value (AUX turns on at roughly 10 dB OBO, albeit slower than expected) compensating the low AUX transconductance (low gain) close to threshold. The DA drain efficiency (Fig. 4, middle) exhibits a neat increase compared the conventional parallel stage class B PA, albeit of course lower than a class C parallel stage. The DA drain efficiency amelioration is up to 10% at 10 dB OBO. The DA amplifier has a lower gain in back-off compared to the class B stage (Fig. 4, right.) but exhibits a milder gain compression, allowing for a better compromise between high efficiency and linearity [8].

REFERENCES

[1] F. H. Raab et. al., "Power amplifiers and transmitters for RF and microwave," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 3, pp. 814-826, March 2002, doi: 10.1109/22.989965.
[2] V. Camarchia et. al., "A K-band GaAs MMIC Doherty power amplifier for point-to-point microwave backhaul applications", *Proc. INMMIC 2014*, 2-4 April 2014, Leuven, Belgium
[3] W. H. Doherty, "A new high efficiency power amplifier for modulated waves", *Proc. Inst. Radio Eng.*, vol. 24, no. 9, pp. 1163–1182, Sep. 1936.
[4] P. M. Asbeck, "Will Doherty continue to rule for 5G?," in *Proc. IEEE MTT-S Int. Microwave Symp. Dig.*, May 2016, pp. 1–4.
[5] Fang J. et. al., "3.5 GHz WiMAX GaN Doherty power amplifier with second harmonic tuning", *Microwave and Optical Technology Letters*, Vol. 54, N. 11, pp 2601–2605, November 2012.
[6] S. Donati Guerrieri, et. al., "A Unified Approach to the Sensitivity and Variability Physics-Based Modeling of Semiconductor Devices Operated in Dynamic Conditions. Part I: Large-signal sensitivity", *IEEE Trans. El. Dev.*, Vol. ED-63, No: 3, pp. 1195–1201, March 2016.
[7] S. Donati Guerrieri, et. al., *IEEE Trans. El. Dev.*, "Concurrent Efficient Evaluation of Small-Change Parameters and Green's Functions for TCAD Device Noise and Variability Analysis", Vol. ED-64, No: 3, pp. 1269–1275, March 2017.
[8] O'Droma M. et. al., "On linearisation of microwave-transmitter solid-state power amplifiers", *Int. J. of RF and Microwave Computer-Aided Engineering* Volume 15, No 5, pp. 491–505, September 2005.

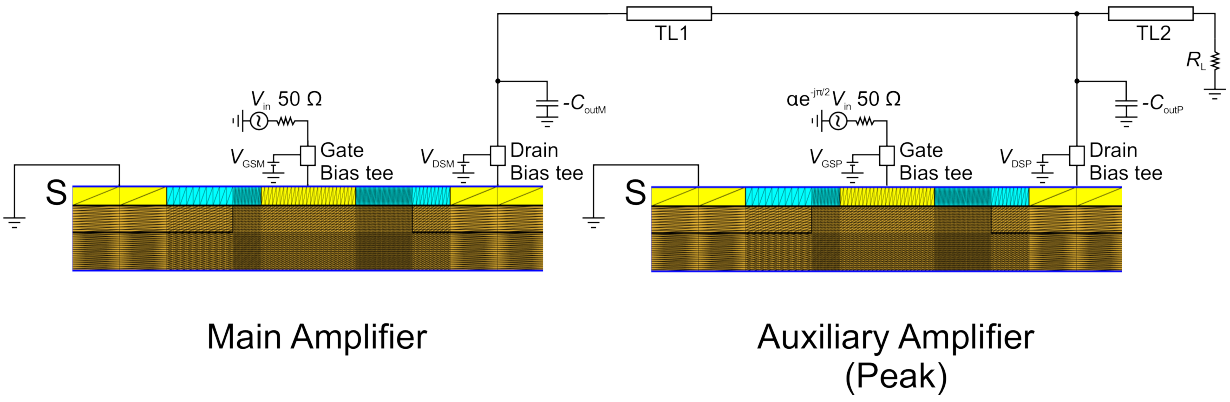


Fig. 1. Schematic representation of the simulated devices and of the embedding circuit within a Doherty amplifier). TL1: pahse 90°, characteristic impedance: $ZC_1 = \sqrt{\alpha} R_{opt}$, where $R_{opt} = \mathrm{Re}(Z_{opt})$ (being Z_{opt} the optimum load impedance) and $\alpha = 1.15$ has been optimized for faster AUX turn on. TL2: pahse 90°, characteristic impedance: $ZC_2 = \sqrt{R_{opt} R_L / 2}$ with $R_L = 50\Omega$.

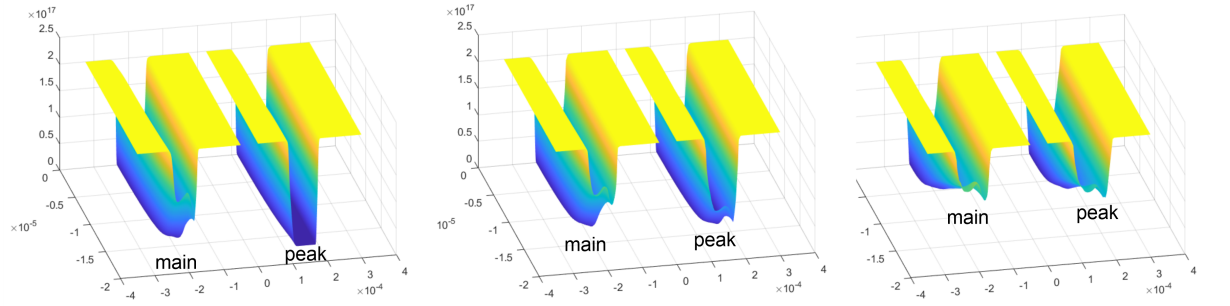


Fig. 2. Electron concentration in the Main and Peak Amplifiers in back-off ($P_{\mathrm{av,M}} = 12.5$ dBm), intermediate power ($P_{\mathrm{av,M}} = 20$ dBm) and peak power ($P_{\mathrm{av,M}} = 26.5$ dBm).

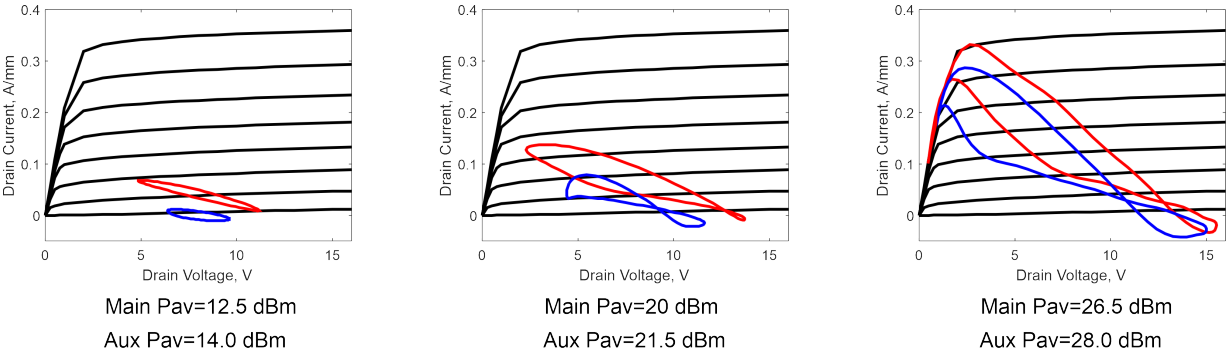


Fig. 3. Dynamic load lines of the Main and Peak Amplifiers in back-off ($P_{\mathrm{av,M}} = 12.5$ dBm), intermediate power ($P_{\mathrm{av,M}} = 20$ dBm) and peak power ($P_{\mathrm{av,M}} = 26.5$ dBm).

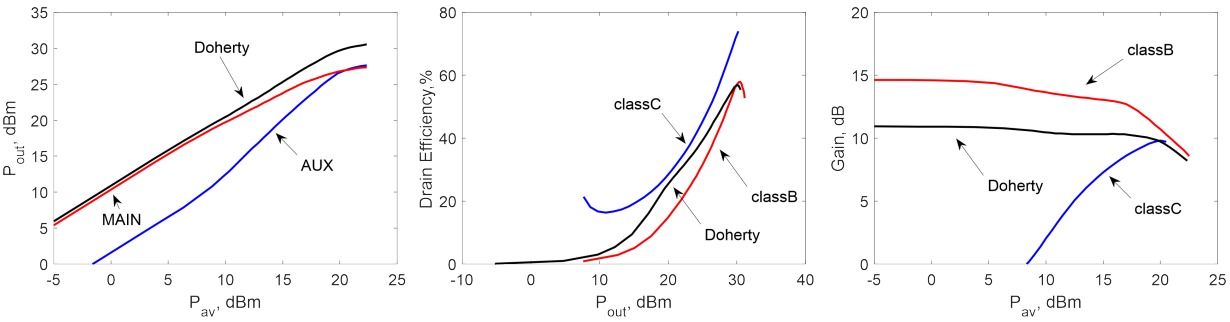


Fig. 4. Output power, drain efficiency and gain of the DA compared to parallel class B and class C stages.

TCAD-Based RF Performance Prediction and Process Optimization of 3D Monolithically Stacked Complementary FET

Shu-Wei Chang,[1] Jia-Hon Chou,[2] Wen-Hsi Lee,[1] Yao-Jen Lee[3] and Darsen D. Lu[4]

[1]Department of Electrical Engineering, National Cheng Kung University, Tainan City, Taiwan 70101
[2]M.S. Degree Program on Nano-IC Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan City, Taiwan 70101
[3]Department of Electrical Engineering, National University of Kaohsiung, Taiwan 81148
[4]Institute of Microelectronics, Department of Electrical Engineering, National Cheng Kung University, Tainan City, Taiwan 70101

Polycrystalline silicon thin-film transistors (poly-TFTs) with low fabrication temperature, high carrier mobility and CMOS compatibility may be optimized for radio frequency (RF) operation to enable system on panel (SoP) or stacked above CMOS devices during back-end-of-line (BEoL) processing for monolithic 3D-IC (M3D) applications. In recent years, three-dimensional structure such as FinFET, nanowire and nanosheet are adopted for optimal device performance. In particular, the on-state current of nanowire and nanosheet devices is enhanced multi-fold with stacked channels, while maintaining excellent gate controllability through their gate-all-around structure. To continue footprint scaling, we have demonstrated a CFET technology which consists of pFETs stacked on top of nFETs, with gate-all-around junctionless nanosheet design, for M3D logic applications [1]. In this study, we further fabricate poly-Si RF devices along with CFET-based logic circuit, realizing heterogeneous integration. Like CFETs, the RF devices also have upper p-type and lower n-type channels. This not only boosts device density but also extends CFET towards diverse applications. However, direct application of CFET towards RF without optimization is inappropriate due to parasitic capacitances between the stacked channels and the extra gate-to-drain capacitance, which lead to poor cutoff frequency (f_T) and maximum oscillation frequency (f_{max}). In this study, we simulate a CFET device which mimics the process and device structure in our previous experimental work with TCAD [2] and analyze its electrical properties, focusing on high frequency performance. First, the transfer characteristics (I_D-V_G) of the CFET model is calibrated to measurements by adjusting the surface scattering model to account for lower mobility in the polysilicon channel. The simulated device characteristics is in good agreement with experimental data (**Fig. 1**). Subsequently, we investigate the impact of tuning individual process parameters, including gate length (L_G), channel width (W), sacrificial oxide thickness (T_S), or the spacing between two silicon channels, release width (W_R), or the lateral-etched width of the sacrificial layer, and metal gate thickness (T_{MG}) (**Fig. 2**), analyzing the impact of each parameter. The outcome is listed in Table I. Take T_S for example, not only f_T but also f_{max} increase with thicker T_S (**Fig. 3**). As the fringe electric field between the two layers becomes less strength with thicker T_S (**Fig. 4**), the capacitive coupling becomes weaker, thus operational frequency is higher. The simulation results shown in **Fig. 5** highlights the fact that G_m increases by 36.9% and gate resistance decreases by 34.6% when T_S increases from 20nm to 70nm. Meanwhile the fringe capacitance between the two channel layers reduces, leading to 21.9% lower gate capacitance (not shown). The overall f_T and f_{max} are improved by 70.5% and 95.6%, respectively. We concluded that a thin oxide layer between the two channels is undesirable for RF. Process-wise, the middle oxide layer should initially be made as thick as possible to prevent the bending of the channel during channel release process, leading to the degradation of device characteristics. A more detailed account of the optimization of the other components (L_G, W_R, W, T_{MG}) will be provided along with the full paper. Finally, we optimize all 5 parameters as part of a performance step-up effort to obtain an optimized CFET structure for RF without adjustment of the process flow. Table II summarizes the optimal CFET structure for RF and compares it to the initial un-optimized design point. f_T and f_{max} are improved by 3.74 and 8.44 times, respectively. Thanks to a flexibility to implement heterogeneous integrated logic, memory, and analog/RF devices on a single chip, we may overcome the technology scaling bottleneck due to physical limitation and the increased interconnect delays that naturally occurs with technology scaling.

1. P. -J. Sung et al., "Fabrication of Vertically Stacked Nanosheet Junctionless Field-Effect Transistors and Applications for the CMOS and CFET Inverters," in IEEE Transactions on Electron Devices, vol. 67, no. 9, pp. 3504-3509, Sept. 2020, doi: 10.1109/TED.2020.3007134.

2. Sentaurus Process, version 2020.09.
3. J. H. Chou. "Process TCAD for RF Performance Step-Up of Three-dimensional Stackable Complementary FET and Improvement Suggestions", MS Thesis, M.S. Degree Program on Nano-IC Engineering, Dept. of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, 2021.

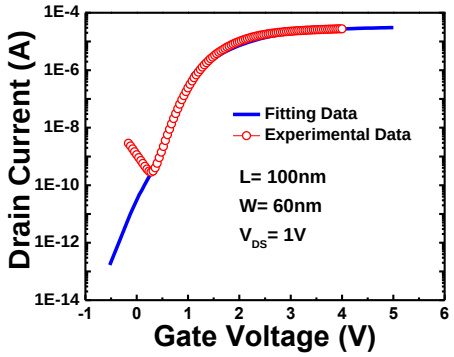


Fig. 1. The I_D-V_G curve simulated using TCAD agree well with experimental data.

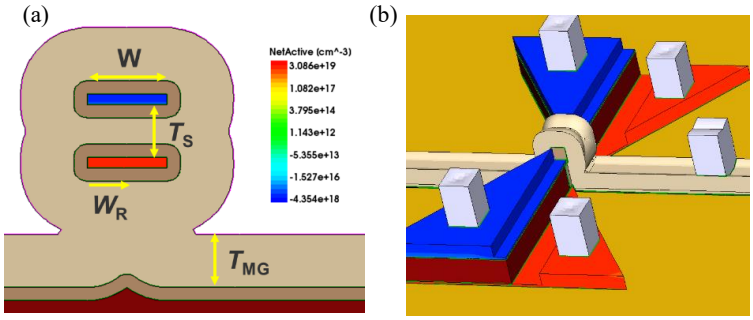


Fig. 2. (a) The cross section and (b) top view of the CFET generated with Sprocess TCAD simulation.

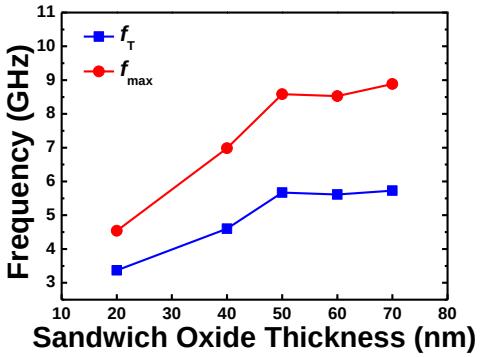


Fig. 3. Cutoff frequency and maximum oscillation frequency versus sacrificial oxide thickness.

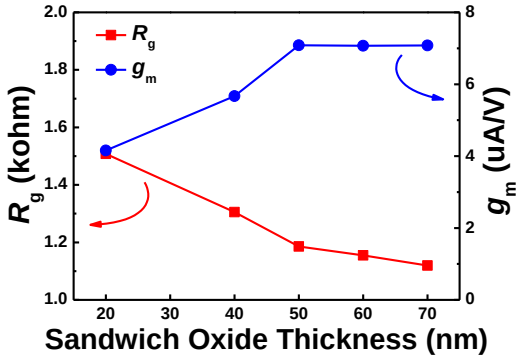


Fig. 5. Relationship of sacrificial oxide thickness with transconductance and gate resistance.

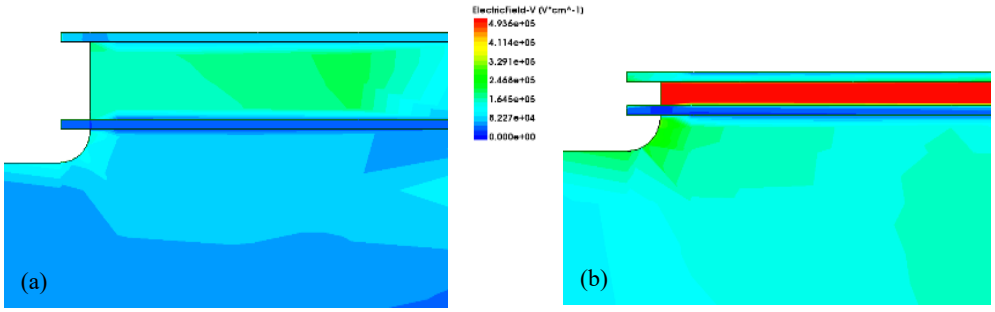


Fig. 4. The electric field distribution of (a) T_S= 70nm and (b) T_S= 20nm.

Table I. Summary of TCAD simulation results of change in f_T and f_{max} after optimizing individual process parameters.

	f_T	f_{max}
L_G (150→60)	+17.3%	+36.5%
W (40→100)	-18%	+18.3%
T_S (20→70)	+70.5%	+95.6%
W_R (30→5)	-29.1%	-49.2%
T_{MG} (100→60)	+10%	+27.5%

Table II. Comparison of RF performance between the calibrated model (Calibration) and optimized model (Optimization).

	Calibration	Optimization
L_G (nm)	100	60
W (nm)	60	60
T_S (nm)	20	60
W_R (nm)	30	30
T_{MG} (nm)	40	100
f_T (GHz)	5.3	19.8
f_{max} (GHz)	6.45	54.41

THz Gain Compression in Nanoscale FinFETs

1st Mathias Pech
Chair for High Frequency Techniques
TU Dortmund
Dortmund, Germany
mathias.pech@tu-dortmund.de

2nd Dirk Schulz
Chair for High Frequency Techniques
TU Dortmund
Dortmund, Germany
dirk2.schulz@tu-dortmund.de

***Abstract*—The gain compression and excitation of higher order harmonics in 3 nm wide on-insulator type FinFETs is investigated. The high computational burden related to the time-resolved analysis needed is dealt with by applying a coupled mode-space approach onto a Quantum Liouville-type Equation. An increase in gain compression and higher order harmonic distortion when decreasing the fin height from 5 to 3 nm is observed.**

***Index Terms*—computational nanotechnology, quantum transport, quantum liouville, mode space, FinFET, gain compression.**

I. Introduction

Current interest in analog and mixed signal circuits is as high as ever, with applications ranging from mixers and general systems on chip (SoC) to specialized architectures for high-performance computing (HPC) that can possibly outperform their digital counterparts [1]. Multigate-FETs with Gate-All-Around- (GAA) and Fin-type gate architectures are especially important due to their excellent channel control and ease of integration into stacked transistor structures [2]. For amplifier and mixer applications, time-resolved quantum transport simulations are necessary to study the nonlinear behavior of these multigate-FETs. Quantum Liouville-type approaches in particular have been demonstrated to be well suited, but have been mostly limited to either stationary or transient low-dimensional devices (eg. [3]). Here, the approach is extended to the three-dimensional regime utilizing a Quantum Liouville-type Equation (QLTE) in order to evaluate gain compression in FinFETs. To further reduce computation time, the transport is effectively projected onto the dominant transport direction. This is done by approximating the density matrix in transport direction with an expansion in terms of the eigensolutions of Schrödinger's equation in the confinement direction, the so-called modes [4]. The resulting method is called mode-space approach, which herein also takes the coupling between different modes into account.

II. Quantum transport in mode-space

An appropriate method based on the solution of a Quantum Liouville type equation (QLTE), which is self-consistently solved along with Poisson's equation is extended to allow an efficient three-dimensional analysis. Starting from Schrödinger's equation in the effective mass- and Hartree approximation, it can be solved assuming a confinement in the yz-plane and a dominant transport in x-direction. This results in modes described by their wave functions ψ_m and corresponding subband energies E_m. The expansion of the wave function Ψ of Schrödinger's equation in terms of the modes ψ_m leads to the ansatz

$$\Psi_k(x,y,z,t) = \sum_m \varphi_m(x,t)\psi_m(y,z,t) \tag{1}$$

with expansion coefficients φ_m. Decoupled Schrödinger equations for each mode ψ_m result, which can be written as

$$\begin{aligned} \imath\hbar\frac{\partial}{\partial t}\varphi_m(x,t) = &\left(-\frac{\hbar^2}{2m_x}\frac{\partial^2}{\partial x^2} + E_m(x,t)\right)\varphi_m(x,t) \\ &- \left(K_{m,1}(x,t)\frac{\partial}{\partial x} + K_{m,2}(x,t)\right)\varphi_m(x,t). \end{aligned} \tag{2}$$

In contrast to [5], coupling between the modes is no longer neglected but included by the coupling terms $K_{m,1}(x,t)$ and $K_{m,2}(x,t)$. The latter leads to the idea of introducing an effective potential $V_m(x,t) = E_m(x,t) - K_{m,2}(x,t)$ for each mode m. After transformation of the density matrix $\rho(x,x')$ onto center-of-mass coordinates (χ,ξ) and introducing the functions

$$\begin{aligned} K^{\pm}_{m,1}(\chi,\xi,t) &= K_{m,1}(\chi+\frac{\xi}{2},t) \pm K_{m,1}(\chi-\frac{\xi}{2},t) \\ B_m(\chi,\xi,t) &= V_m(\chi+\frac{\xi}{2},t) - V_m(\chi-\frac{\xi}{2},t), \end{aligned} \tag{3}$$

the QLTE can be derived in the so-called mode-space as:

$$\begin{aligned} \frac{\partial}{\partial t}f_m(\chi,k,t) = \Bigg[&\Big(-\frac{\hbar k}{m_x} + \frac{\imath}{4\hbar}\int\frac{dk'}{2\pi}\tilde{K}^{-}_{m,1}(\chi,k'',t)\Big)\frac{\partial}{\partial\chi} \\ &+ \frac{\imath}{2\hbar}\int_{-\infty}^{\infty}\frac{dk'}{2\pi}\tilde{K}^{+}_{m,1}(\chi,k'',t)\cdot(\imath k') \\ &+ \frac{1}{\imath\hbar}\int_{-\infty}^{\infty}\frac{dk'}{2\pi}\tilde{B}_m(\chi,k'',t)\Bigg] f_m(\chi,k',t). \end{aligned} \tag{4}$$

Here, $k'' = k - k'$ has been introduced and the tilde denotes integral kernels regarding $K_1^{\pm}$ and B. The coordinate χ indicates the transport direction.

III. Investigation of Gain Compression in FinFETs

Key parameters regarding the size of the FinFET in question are shown in Fig. 1. The undoped $In_{0.53}Ga_{0.47}As$ channel is 10 nm long, as is the gate contact. For the oxide and gate

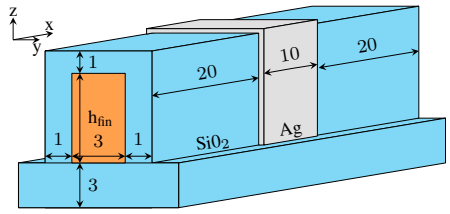


Fig. 1. Schematic of the FinFET with all dimensions in nm.

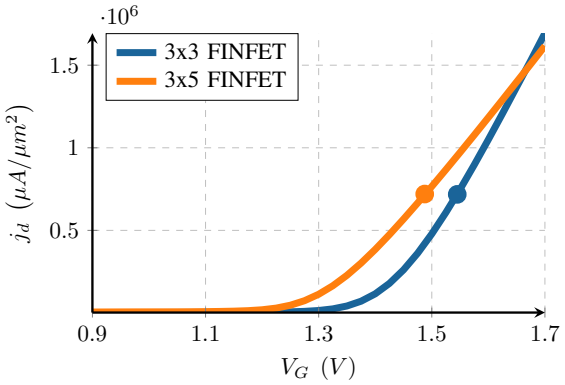


Fig. 2. Transfer curves of the FinFETs with marks indicating operating points.

material SiO_2 and Ag are chosen, respectively. The source and drain regions on either end of the device are n-doped with $N_s = N_d = 2 \cdot 10^{19}\ cm^{-3}$. The QLTE is discretized utilizing a finite volume scheme as described in [3]. Inflow boundary conditions to model carrier transport in and out of the device are adopted [6] and a complex absorbing potential is added to avoid nonphysical solutions [7]. A harmonic gate voltage of $V_G = V_0 + V_A \cdot \sin(2\pi \cdot f_0 \cdot t)$ with the center frequency $f_0 = 250$ GHz is applied. The operating points V_0 are indicated in Fig. 2. Values from 0 V to 0.3 V are chosen for the amplitude V_A with increments of $\Delta V_A = 0.025$ V. Two different FinFETs with a fixed fin width of 3 nm and fin heights of 3 and 5 nm are considered for demonstration purposes.

As it is evident from Fig. 3, compression during the positive half-wave takes place sooner in the 3x3 nm FinFET when compared to the 3x5nm FinFET, thus leading to an increase in distortion. The drain-end current density can also be analyzed in terms of the spectral components at $f_0 = 250$ GHz and the harmonics $f_1 = 500$ GHz and $f_2 = 750$ GHz. When increasing the channel dimensions from 3x3 nm to 3x5 nm, the latter shows less excitation of higher order harmonics and higher gain, as it can be seen from Fig. 4.

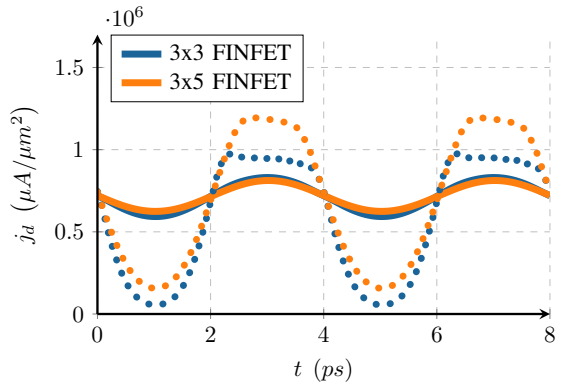


Fig. 3. Time-dependent drain-end current density at $V_A = 0.025V$ (—) and $V_A = 0.175$ (· · ·).

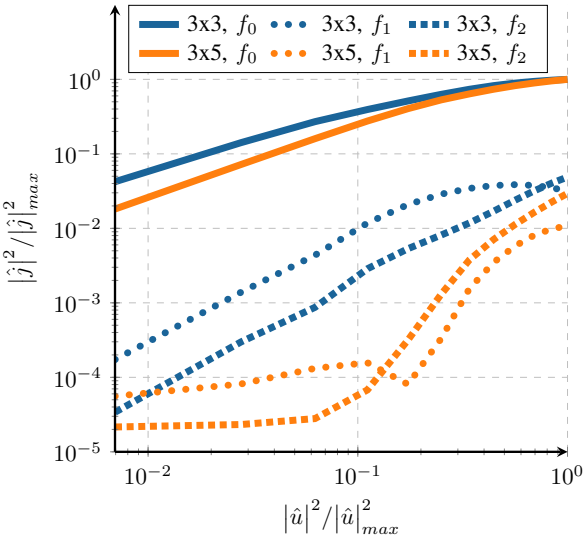


Fig. 4. Normalized and squared Fourier Coefficients of input (gate voltage) and output (drain-end current density) for the center frequency f_0 and the higher order harmonics f_1 and f_2.

References

[1] S. Köppel, B. Ulmann, L. Heimann, and D. Killat, "Using analog computers in today's largest computational challenges," *Advances in Radio Science*, vol. 19, pp. 105–116, 2021. [Online]. Available: https://ars.copernicus.org/articles/19/105/2021/

[2] N. Jha and D. Chen, *Nanoelectronic circuit design*, 2011.

[3] L. Schulz and D. Schulz, "Formulation of a phase space exponential operator for the wigner transport equation accounting for the spatial variation of the effective mass," *Journal of Computational Electronics*, vol. 19, 2020.

[4] Z. Ren *et al.*, "nanomos 2.5: A two-dimensional simulator for quantum transport in double-gate mosfets," *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1914–1925, 2003.

[5] L. Schulz and D. Schulz, "Time-resolved mode space based quantum-liouville type equations applied onto dgfets," in *2020 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2020, pp. 331–334.

[6] W. R. Frensley, "Boundary conditions for open quantum systems driven far from equilibrium," *Rev. Mod. Phys.*, vol. 62, pp. 745–791, Jul 1990. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.62.745

[7] L. Schulz and D. Schulz, "Complex absorbing potential formalism accounting for open boundary conditions within the wigner transport equation," *IEEE Transactions on Nanotechnology*, vol. 18, pp. 830–838, 2019.

SISPAD 2022, September 6-8, 2022, Granada, Spain

Ab initio study of electron mobility in V_2O_5 via polaron hopping

R. Defrance[1, 2], B. Sklénard[1], M. Guillaumont[2], J. Li[1], M. Freyss[3]
[1] Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France.
[2] LYNRED, 364 Avenue de Valence, 38113 Veurey-Voroize, France.
[3] CEA, DES, IRESNE, DEC, Cadarache, 13108 Saint-Paul-lez-Durance, France.

Abstract: *We investigated the polaron transport in V_2O_5 using density functional theory (DFT)+U. The Bond Distortion Method (BDM) is employed to stabilize an excess electron as a self-trapping polaron in V_2O_5. The polaron hopping is evaluated by Landau-Zener equation with parameters extracted from the DFT ground state energies on the Linear Interpolation (LI) reaction coordinates using a model Hamiltonian. Electron mobility contributed by polaron hopping is obtained from Einstein relation and compared to experimental data. The influence of the Hubbard U value in DFT+U on both the polaron static and transport properties are reported.*

Keywords: *Polaron, V_2O_5, DFT+U, trapping, hopping, mobility.*

1) Introduction

Vanadium pentoxide (V_2O_5) is a Transition Metal Oxide (TMO) with several interesting properties that can be used in fields such as energy storage [1]. It is therefore important to study the electron transport in crystalline V_2O_5 which is known to be led by small polarons hopping [2]. For such materials with strongly correlated *d*-orbitals, standard local and semi-local DFT calculations fail to predict the electronic properties. The main reason of this failure is the self-interaction error which can be circumvented by introducing an on-site Hubbard U correction for the 3d V electrons in the so-called DFT+U method. The value of U is usually determined empirically (e.g. adjusted to reproduce properties such as the band gap) even though it can be calculated with *ab initio* methods [3]. In this work, we investigate the impact of U on polaron properties, from its localization to its contribution to electron mobility in V_2O_5.

2) Methods

All DFT+U calculations presented in this work are carried out using the PAW method implemented in the VASP code and the Perdew-Burke-Ernzerhof (PBE) [4] exchange-correlation functional. To stabilize the polaron, we employ a 168-atom supercell of orthorhombic V_2O_5 (Fig. 1) and we apply the Bond Distortion Method (BDM) with an initial distortion of 15% around one of the V atoms to localize the excess electron before relaxation.

We use the Linear Interpolation (LI) method to simulate the polaron jump. The coordinates of all the atoms along the hopping path are approximated as: $\boldsymbol{R} = x\boldsymbol{R_A} + (1-x)\boldsymbol{R_B}$, where $\mathbf{R_A}$ and $\mathbf{R_B}$ are the initial and final coordinates and x varies from 0 to 1. Our method to extract key parameters of the transport is a fit to a two states model Hamiltonian:

$$\boldsymbol{H} = \begin{bmatrix} H_{AA} & H_{AB} \\ H_{BA} & H_{BB} \end{bmatrix} (1)$$

with $H_{AA} = \lambda_A r^2 + \boldsymbol{E_{Anhar}}(r^3)$, $H_{BB} = \Delta G^0 + \lambda_B(1-r)^2 + \boldsymbol{E_{Anhar}}(r^3)$ and $H_{AB} = H_{BA}$ the coupling constant. The harmonic contribution in the expression of H_{AA} and H_{BB} follows the Marcus theory [5] to describe the polaron transition from state A to B as illustrated in Fig. 2. We also include an anharmonic contribution to improve the fit. Electronic mobilities were calculated based on the hopping rate obatined from the Landau-Zener equation [6]:

$$k_{ET} = \kappa_{el} \nu_{eff} \Gamma exp\left(-\frac{E_a}{k_B T}\right) (2)$$

with ν_{eff} the effective phonon frequency, $\kappa_{el} = \frac{2P_{LZ}}{(1+P_{LZ})}$, $P_{LZ} = 1 - \exp\left(-\frac{\frac{3}{\pi^2}|H_{AB}|^2}{h\nu_{eff}\sqrt{\lambda_A k_B T}}\right)$, $\Gamma = 1$ the nuclear tunneling factor, E_a the activation energy and ΔG^0 is the energy difference between state A and B. The polaron mobility is then evaluated from the Einstein relation:

$$\mu = \frac{|q|\,D}{k_B T} \; with \; D = r^2 n_{neigh} k_{ET} \; (3)$$

where q is the charge of the carrier, r is the distance between the two sites and n_{neigh} is the number of neighbouring sites.

3) Stabilization and hopping of polaron.

We first stabilized the polaron with the BDM and obtained results close to the literature [7]-[8]. However, we show in Table 1 that the value of U has a significant impact on the polaron properties. An increase of U leads to a decrease of the polaron formation energy (E_{POL}) as defined in [9], and an increase of the magnetic moment (Magn.) and of the lattice distortion around the polaron (average bond distortion of the 6 nearest neighbours), enhancing the stabilization of the polaron.

Then, we simulate the hopping of the polaron along the [100], [010] and [001] directions using the LI method for U=3.25 eV. The calculated barriers and the energy profile are anisotropic as shown in Fig. 3 and Fig. 4. As we fit the Hamiltonian (Eq.1) to the DFT results, the shape of the energy curve directly impacts the value of the extracted parameters reported in Table 2. By looking at the ratio $\frac{\lambda_A}{H_{AB}}$, we can say that the hopping in the [100] direction is adiabatic while the hopping in the two other directions is nonadiabatic [5], [10]. We can confirm this observation by looking at the charge transfer along the jump displayed on Fig. 5. In the [100] direction the charge follows the displacement (adiabatic behaviour) while in the [010] and [001] ones the charge is transferred from site A to B near the saddle point (nonadiabatic behaviour).

Moreover, we show on Fig. 5 that as U increases, the barrier also increases. This behaviour is consistent with the influence of U on polaron properties shown in Table 1. Furthermore, we also notice in Fig. 5 and Fig. 6 that the increase of U changes the energy profile of the barrier and the hopping dynamics. As U increases, the [100] hopping becomes more and more nonadiabatic as the barrier looks more sharp and the localized electron follows less the reaction coordinate.

4) Polaron mobilities in crystalline V_2O_5

To compute the polaron mobility, we consider only nearest neighbours for the hopping and we use $\nu_{eff} = 10^{13} Hz$ which is a typical value for TMO[11]. The reported experimental mobility values measured at room temperature in V_2O_5 are scattered and range from $1.84 \times 10^{-2} cm^2 V^{-1} s^{-1}$ [12] to $4.75 cm^2 V^{-1} s^{-1}$ [13]. The calculated mobility depends on the hopping direction and the value of U as shown in table 2. The mobility ($\mu = 5.5 \times 10^{-2} cm^2 V^{-1} s^{-1}$) along the [100] direction, computed with

U=3.25eV, is close to the value reported in [12]. Other directions and higher values of U give smaller mobilities which means that the hopping along the [100] direction is dominant. Results of Table 2 also illustrate the strong dependence on the U value as the mobility is shifted by several orders of magnitude with a change of only 1 eV of U.

5) Conclusion

We investigated the static and the transport properties of polaron in monocrystalline V_2O_5 starting from *ab initio* DFT+U method. We demonstrated the extraction of polaron hopping parameters by mapping the DFT ground state energies on the reaction coordinates to a model Hamiltonian. With our methodology we emphasise the sensitivity to the choice of Hubbard U, from the polaron self-trapping to its mobility. With U=3.25 eV (which is used in literature for V_2O_5 [14]), the computed mobility is in good agreement with some of the experimental data reported in the literature. However, it is difficult to attribute experimentally one exact value of mobility for monocrystalline V_2O_5, because of its sensibility to defects (such as oxygen vacancies) during the fabrication process.

[1] M. Sathiya et al, *J. Am. Chem. Soc.*, vol. 133, nº 40, p. 16291-16299, oct. 2011.
[2] L. Murawski et al, *Journal of Non-Crystalline Solids*, vol. 89, nº 1-2, p. 98-106, jan. 1987.
[3] B.-C. Shih et al, *Phys. Rev. B*, vol. 86, nº 16, p. 165124, oct. 2012.
[4] J. P. Perdew et al, *Phys. Rev. Lett.*, vol. 77, nº 18, p. 3865-3868, oct. 1996.
[5] R. A. Marcus, *Rev. Mod. Phys.*, vol. 65, nº 3, p. 599-610, july. 1993.
[6] F. Wu et al, *J. Mater. Chem. A*, vol. 6, nº 41, p. 20025-20036, 2018.
[7] L. Ngamwongwan et al, *Phys. Chem. Chem. Phys.*, vol. 23, nº 19, p. 11374-11387, 2021.
[8] P. Watthaisong et al, *RSC Adv.*, vol. 9, nº 34, p. 19483-19494, 2019.
[9] Reticcioli M. et al (2019). In: Andreoni W., Yip S. (eds) Handbook of Materials Modeling. Springer, Cham.
[10] N. A. Deskins et al, *Phys. Rev. B*, vol. 75, nº 19, p. 195212, may 2007.
[11] S. Pal et al, *phys. stat. sol. (b)*, vol. 237, nº 2, p. 513-522, june 2003.
[12] J. C. Badot et al, *J. Mater. Chem.*, vol. 14, nº 23, p. 3411, 2004.
[13] J. Haemers et al, *Phys. Stat. Sol. (a)*, vol. 20, nº 1, p. 381-386, nov. 1973.
[14] N. J. Szymanski et al, *Computational Materials Science*, vol. 146, p. 310-318, apr. 2018.

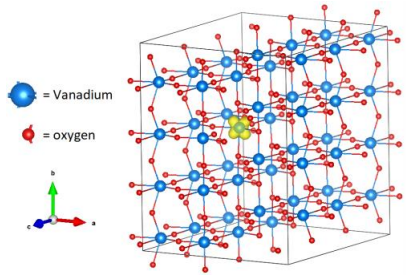


Figure 1 : Relaxed supercell of orthorhombic V_2O_5 with the charge density isosurface of the polaron (yellow) localized on a V atom.

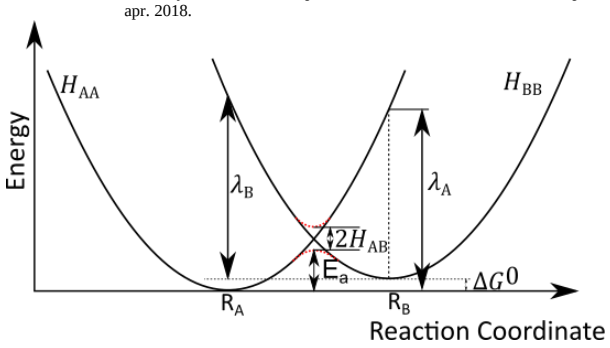


Figure 2 : Global description of the Marcus theory for a two-state transfer.

U(eV)	E_{POL}(eV)	Magn.(μB)	Dist.(%)
3	0.10	0.992	3.74
3.25	0.25	1.022	3.73
3.5	0.25	1.045	4.07
4	0.40	1.082	4.33
4.5	0.42	1.112	4.54
5	0.70	1.138	4.73

Table 2 : Properties of the polaron depending on the U value: the polaron formation energy (E_{POL}), the magnetic moment on the Vanadium atom with the localized electron (Magn.) and the crystal distortion around the polaron (Dist.).

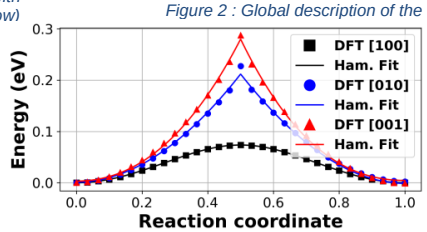


Figure 3 : Polaron hopping energiy barrier calculated with the LI method along 3 directions with U=3.25eV. Markers are DFT data and lines are fit to the Hamiltonian.

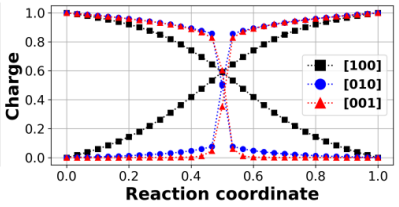


Figure 4 : Charge transfer during the hopping along the three main directions with the LI method, dash lines are guide to the eye. Markers are magnetic moment normalized from 0 to 1.

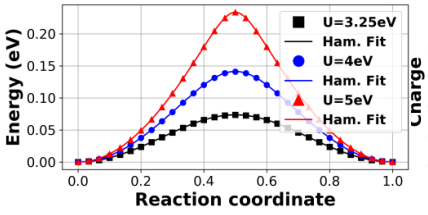


Figure 5 : Polaron hopping energy barrier calculated with the LI method with U=3.25eV (black), U=4eV (blue) and U=5eV (red). Markers are DFT data and lines are fit to the Hamiltonian.

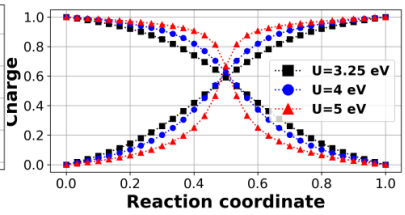


Figure 6 : Charge transfer during the hop along the [100] direction with the LI method with U=3.25eV(black), U=4eV (blue) and U=5eV (red).

U(eV)	Direction	Linear Interpolation				
		E_a^{DFT}(eV)	E_a^{Ham}(eV)	H_{AB}(eV)	k_{ET}(300K)(Hz)	μ(300K)($cm^2V^{-1}s^{-1}$)
3.25	[100]	0.07	0.07	0.14	5.82×10^{11}	5.5×10^{-2}
	[010]	0.23	0.22	5.9×10^{-5}	1.64×10^{4}	1.6×10^{-9}
	[001]	0.29	0.28	5.39×10^{-7}	1.08×10^{-1}	1.4×10^{-14}
4	[100]	0.14	0.14	0.11	4.25×10^{10}	4×10^{-3}
	[010]	0.26	0.26	4.2×10^{-6}	1.28×10^{1}	1.3×10^{-12}
	[001]	0.30	0.30	1×10^{-5}	1.34×10^{1}	1.7×10^{-12}
5	[100]	0.23	0.23	0.06	1.14×10^{9}	1.1×10^{-4}
	[010]	0.30	0.30	1.03×10^{-3}	1.34×10^{5}	1.3×10^{-8}
	[001]	0.34	0.34	4.36×10^{-3}	4.63×10^{5}	5.8×10^{-8}

Table 1 : Transport parameters extracted from the fit to the Hamiltonian depending on the value of U and on the direction: $\Delta G^*$is the barrier energy calculated with DFT+U, E_a is the activation energy of the Landau-Zener model, H_{AB} is the coupling constant, k_{ET} is the transfer rate of Landau-Zener calculated at T=300K and μ is the mobility calculated at 300K with k_{ET} and the Einstein relation.

Efficient Atomistic Simulations of Lateral Heterostructure Devices with Metal Contacts

Mincheol Shin and Kanghyun Joo

School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Rep. of Korea
E-mail: mshin@kaist.ac.kr

Abstract

In this work we present a highly efficient method to perform quantum transport simulations on atomistic devices with metal contacts. In particular, we consider lateral heterostructures of silicide/semi-conductor/silicide and metal/semi-metal/metal which are constructed by the first-principles density functional theory (DFT) method. We show that large-sized heterostructure DFT Hamiltonian can be effectively reduced, while not losing the accuracy in a practical sense, enabling highly efficient calculation of the electrical transport properties of the devices based on the non-equilibrium Green's function (NEGF) method.

Introduction

As the role of the interfaces, junctions and contacts is increasingly important in ultra-scaled devices, it has become necessary to include them as integral parts of simulated device. Although the empirical tight binding (ETB) method combined with the NEGF method is the state-of-the-art atomistic device simulation methodology, it has serious limitations in dealing with heterogeneous structures consisting of different materials. The parameter-free DFT method which is naturally suited for the problem is therefore called for.

A major obstacle of adopting DFT method in device simulations is the computational burden of handling large-sized DFT Hamiltonian. This must be overcome if the DFT Hamiltonian is to be actively used in realistic device simulations in place of ETB. In this regard, the Hamiltonian size reduction method is attractive because the computational time and resources can be reduced by a few orders of magnitude, while the errors in the charge density and current are kept within a few percents or less. The reduction method for DFT Hamiltonian has been developed for homogeneous structures [1,2] and recently extended to treat heterostructures [3]. In this work, we further extend the application of the latter method to heterogeneous structures with metallic contacts, paving the way for realistic simulations of future generation logic and memory devices at the atomistic scale.

Simulation Approach

A generic lateral heterostructure device consisting of three different materials, represented by unit cells A, B, and C, is shown in Fig. 1 (a). Between two materials there are junction cells J_i that represent spatially transient behavior in terms of the atom species and their positions. We form a supercell that consists of minimal number of cells with which the device in Fig. 1 (a) can be constructed. See Fig. 1 (b). We relax the supercell by the DFT method and extract its Hamiltonian which is a block matrix (Fig. 1 (c)). We then perform the Hamiltonian reduction as follows.

The basic idea of the Hamiltonian size reduction method is to unitarily transform the full Hamiltonian to a smaller sized matrix within an energy window of interest where the charge transport takes place. This is possible by constructing a transformation matrix U consisting of a few selected Bloch states within the energy window and removing any unphysical states arising from the transformation [1]. For heterostructures, it is not feasible to obtain the Bloch states of each component cell as some cells like junction cells cannot be repeated periodically. Instead, we solve for the Bloch states of the supercell of Fig. 1 (b) and form a matrix $\mathcal{U}_{ND}$ consisting of the supercell Bloch states. We then make an initial diagonal matrix $\mathcal{U}$ by placing sub-blocks of $\mathcal{U}_{ND}$ to the diagonal entries as shown in Fig. 1 (d). Then we remove the unphysical states on a block-by-block basis so as to retain the block diagonal form of $\mathcal{U}$ [3]. If all the unphysical branches are cleared, we obtain the effective supercell each block of which is reduced in its Hamiltonian size (Fig. 1 (d)). With the effective heterostructure Hamiltonian, we reconstruct the device and apply the NEGF method to calculate the charge density and current (Fig. 1 (e)).

Results

We applied the method described above to calculate the electrical properties of silicide/semiconductor and metal/semi-metal devices. For the former, NiSi/Si/NiSi Schottky barrier structure in the nanowire geometry (Fig. 2 (a)) is considered where NiSi silicide is assumed a crystal phase and shows metallic behavior. The supercell consists of 10 blocks of total 533 atoms, where block 1 and block 6 are NiSi and Si unit cells, respectively, and blocks 3, 4 and 8, 9 are the junction cells (see the caption of Fig. 2 for blocks 10, 2 and 5, 7). The periodic boundary condition was applied to the supercell. We used the SIESTA package to relax the supercell until the maximum force on any of the atoms becomes less than 0.04 eV/Å. A standard GGA-PBE functional with the DZP pseudoatomic orbital basis set was used for the DFT method. In the junction cells, the atoms are seen unorderly positioned as the stress due to the lattice mismatch is applied in the regions. Fig. 2 (b) shows

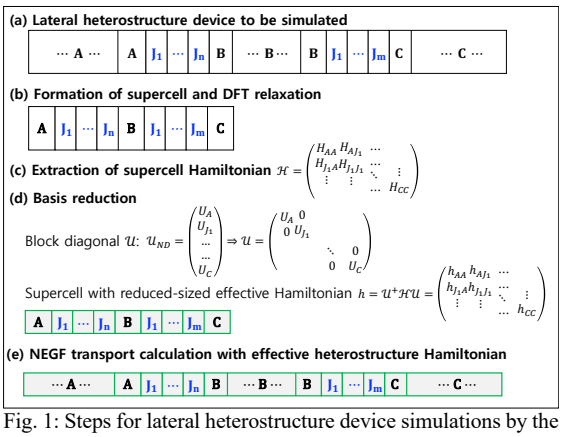


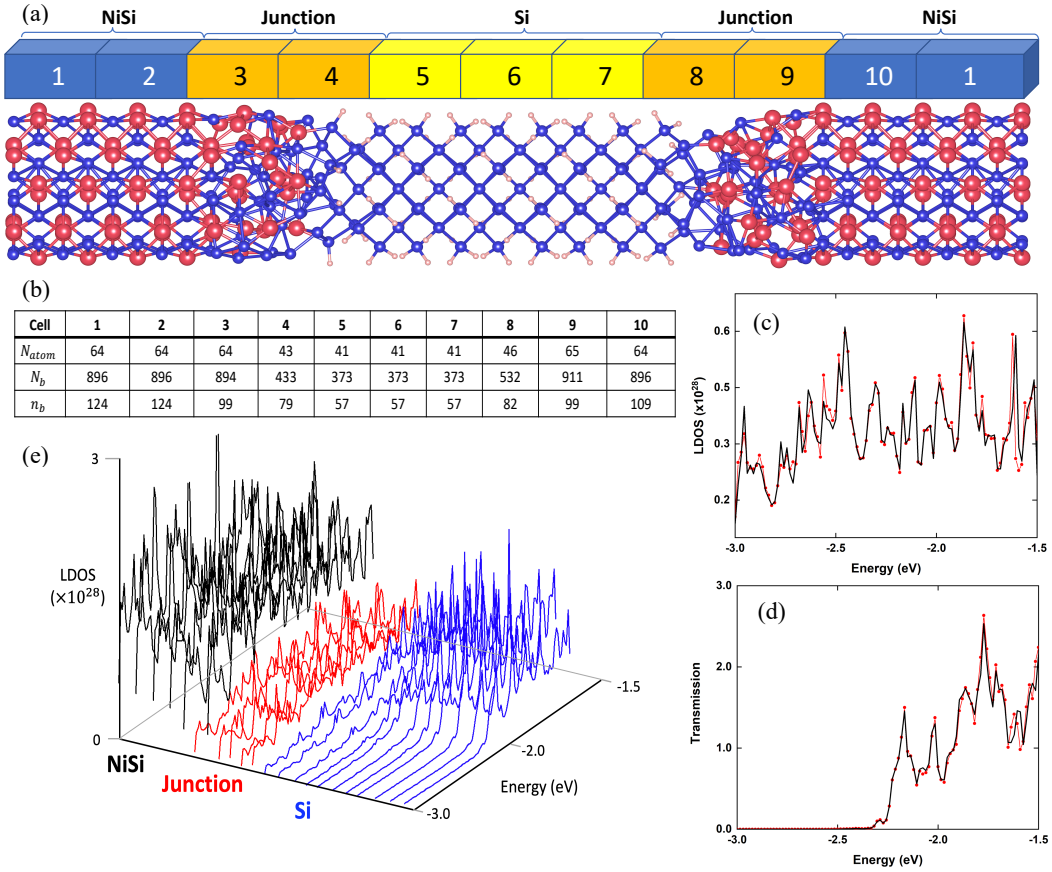
Fig. 1: Steps for lateral heterostructure device simulations by the heterostructure Hamiltonian reduction method. See the text.

the result of the heterostructure Hamiltonian reduction in the energy window of -3.0 eV to -1.5 eV. The Hamiltonian size is reduced by about 15 %. Figs. 2 (c) and (d) show that the local density of states (LDOS) and transmission are excellently reproduced by using the reduced-sized effective Hamiltonian, while the computation speed is more than 2 orders faster. In Fig. 2 (e), the interface energy states are clearly seen at the silicide/semiconductor junction.

Fig. 3 shows TiN/Sb/TiN heterostructure where a-few-layer Sb semi-metal is contacted with TiN metal, with a potential application as a next-generation phase-change memory in mind. The supercell with total of 580 atoms are similarly handled to yield the reduced-sized effective Hamiltonian, which excellently reproduces the transmission in the energy window of 1.5 eV wide (Fig. 3 (b)).

References

[1] M. Shin et al., J. Appl. Phys., 119, 154505 (2016).
[2] M. Shin et al., IEDM 2017.
[3] M. Shin, J. Appl. Phys., 130, 104303 (2021



Cell	1	2	3	4	5	6	7	8	9	10
N_{atom}	64	64	64	43	41	41	41	46	65	64
N_b	896	896	894	433	373	373	373	532	911	896
n_b	124	124	99	79	57	57	57	82	99	109

Fig. 2. (a) NiSi/Si/NiSi supercell consisting of 10 blocks. The supercell in the nanowire geometry is 8.2 nm long and 1.0 nm in diameter. Red and blue balls represent Ni and Si atoms, respectively, and small empty balls hydrogen passivation atoms. Blocks 10, 2, and 5, 7 are the junction cells according to Fig. 1 but may as well be called the buffer cells as they are only slightly different from the unit cells as their colors indicate in the upper panel. (b) Table of number of atoms (N_{atom}), size of full Hamiltonian (N_b) and size of the reduced Hamiltonian (n_b) of each block. (c) LDOS at the left junction and (d) transmission calculated by the full Hamiltonian (black lines) and reduced-sized Hamiltonian (red dots with thin lines). (e) LDOS profile around the left junction. The energy states in the junction region are highlighted by the red color.

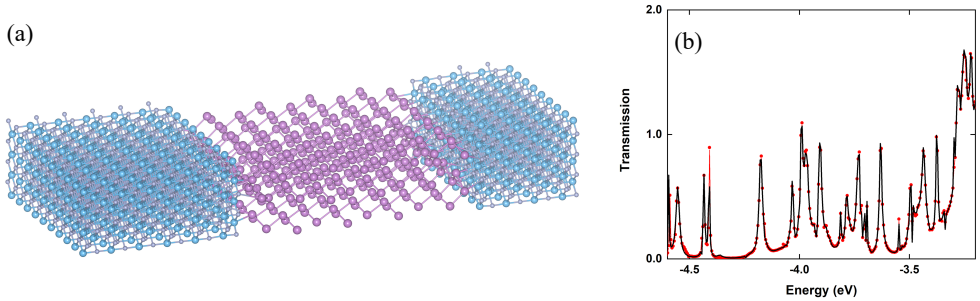


Fig. 3: (a) TiN/Sb structure in the thin film geometry. The supercell is 7.5 nm long,1.0 nm thick and periodic in the perpendicular direction. (b) Transmission calculated by the full Hamiltonian (black lines) and reduced-sized Hamiltonian (red dots with thin lines).

Impact of random alloy fluctuations on carrier transport in (In,Ga)N quantum well systems: Linking atomistic tight-binding models to drift-diffusion

M. O'Donovan[1,2],*, P. Farrell[3], T. Streckenbach[3], T. Koprucki[3] and S. Schulz[2,1]
[1]*Tyndall National Institute, University College Cork, Cork, T12 R5CP, Ireland*
[2]*Department of Physics, University College Cork, Cork, T12 YN60, Ireland*
[3]*Weierstrass Institute (WIAS), Mohrenstr. 39, 10117 Berlin, Germany*
*email: michael.odonovan@tyndall.ie

The semiconductors InN, GaN and their alloy (In,Ga)N have attracted significant attention for optoelectronic device applications in the visible spectral range [1]. Unlike other III-V materials, the electronic and optical properties of (In,Ga)N heterostructures are strongly impacted by alloy fluctuation-induced carrier localization effects [2]. Recent studies have focused on the impact carrier localization has on charge carrier transport in (In,Ga)N based light emitting diodes (LEDs) [3].

In this work we study carrier transport in (In,Ga)N/GaN quantum well (QW) systems by connecting an atomistic tight-binding (TB) model to a 3D continuum-based drift-diffusion framework utilizing ddfermi [4,5]. A schematic of the workflow is shown in Figure 1. Starting from atomistic tight-binding, we construct a local Hamiltonian at each lattice site which is used to extract an (alloy disordered) energy landscape describing the valence band and conduction band confining potentials [6]. The energy landscape generated on an atomistic level accounts for (local) strain and polarization effects. The obtained (locally fluctuating) band edges are placed on an atomistic finite volume mesh, which is then embedded within a sparse device mesh describing the device where alloy fluctuations are not critical (e.g. (*p*- or *n*-) doped GaN regions). In doing so we can simulate full devices while maintaining a feasible numerical burden. Localization landscape theory has been employed to account for quantum corrections via an effective confining potential [7].

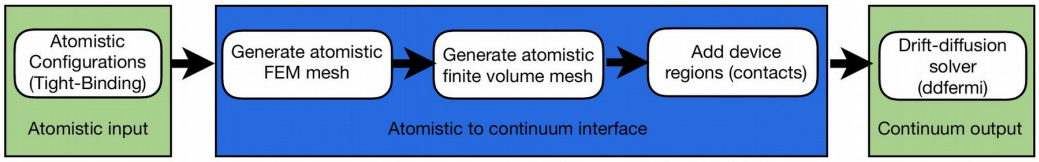


Figure 1: Schematic of workflow to connect the atomistic tight-binding framework to the continuum drift-diffusion solver (ddfermi).

To investigate the impact of alloy fluctuations on carrier transport, random alloy calculations are compared to a virtual crystal approximation (VCA). In a VCA any spatial fluctuations in alloy composition in the QW are neglected, and piecewise constant effective material parameters are used within this region; this is a 3D analogy of the 1D simulations available in many commercial software packages. We use a 3D model in order to compare with random alloy calculations, where a description of the system in-plane is required to capture alloy disorder. To separate the impact of alloy fluctuations on carrier transport from other effects (e.g., recombination) the framework is first applied to uni-polar systems for both electrons and holes by embedding the QWs in *n*- and *p*-doped GaN barriers, respectively.

Our results show that in the case of electrons, a VCA gives a significantly lower current density than the random alloy calculations at fixed bias [4]. Including alloy fluctuations leads to a softening of the potential barriers at the QW interfaces which is not captured in a VCA. Quantum corrections included via localization landscape theory result in a further softening of this barrier which leads to an increase in current in both the VCA and random alloy calculations.

Holes show contrasting results: the current density of a VCA *exceeds* the current density in a random alloy calculation [8]. This is due to the high hole effective mass, which leads to strong alloy-disorder-induced carrier localization within the QW region when including alloy fluctuations. This is visualized in Figures 2 and 3 where the valence band edge and hole charge densities are shown for random alloy calculations excluding (purple) and including (green) quantum corrections; VCA results are also presented (black, dashed). Here the impact of the alloy fluctuations is visible in the valence band edge leading to an increased hole density within the QW compared to the VCA; the coupling of the valence band edge, Fermi levels and charge densities with the device potential leads to a depletion of holes in the GaN barrier material. In turn, this results in an increased resistance in the barrier due to fluctuations in the QW. The quantum corrections have the impact of reducing the strong fluctuations in the QW band edges, however, a depletion of charge carriers in the barrier still occurs resulting in a reduced current compared to VCA.

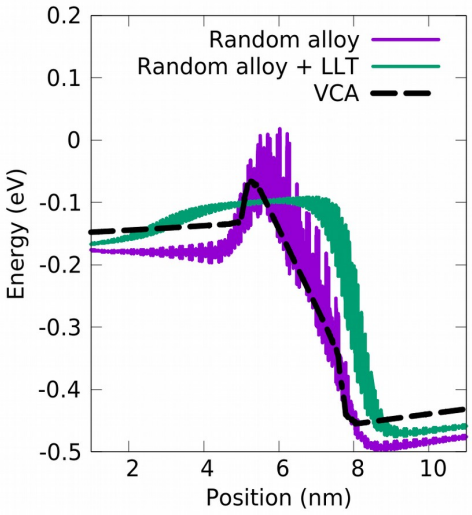


Figure 2: Valence band edge profile along the transport direction of a 3 nm single quantum well (QW) embedded in p-doped GaN barriers at equilibrium. Results are shown for a random alloy calculation excluding quantum corrections (purple), a random alloy calculation including quantum corrections via localization landscape theory (LLT) (green) and a virtual crystal approximation (black, dashed).

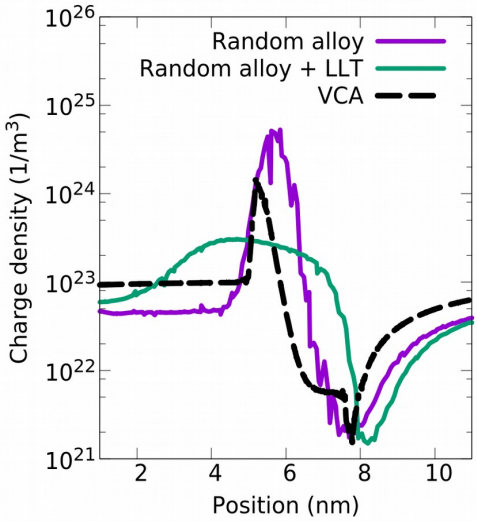


Figure 3: Hole charge density averaged over each atomic plane in a single quantum well (QW) embedded in p-doped GaN barrier at a bias of 1.0 V. Results are shown for a random alloy calculation excluding quantum corrections (purple), a random alloy calculation including quantum corrections via localization landscape theory (LLT) (green) and a virtual crystal approximation (black, dashed).

Equipped with this framework we extend our study to an LED *p-i-n* junction. We will present results on the impact of alloy fluctuations on charge carrier distribution and recombination processes across an (In,Ga)N/GaN multi-QW system.

References

[1] C. J. Humphreys, MRS Bulletin **33**, 459 (2008)
[2] S. Schulz *et al.*, Phys. Rev. B. **91,** 035439 (2015)
[3] C.-K. Li *et al.*, Phys. Rev. B. **95,** 144206 (2017)
[4] M. O'Donovan *et al.*, J. Appl. Phys. **130**, 065702 (2021)
[5] D. H. Doan *et al.*, http://doi.org/10.20347/WIAS.SOFTWARE.DDFERMI (2016)
[6] D. Chaudhuri *et al.*, J. Appl. Phys. 129, 073104 (2021)
[7] M. Filoche *et al.*, Phys. Rev. B. **95,** 144204 (2017)
[8] M. O'Donovan *et al.*, preprint at https://doi.org/10.48550/arXiv.2111.01644 (2021)

Robust Cryogenic *Ab-initio* Quantum Transport Simulation for L_G=10nm Nanowire

Tom Jiao and Hiu Yung Wong*
Electrical Engineering, San Jose State University, CA, USA
*hiuyung.wong@sjsu.edu

Abstract – Motivation and Achievement

Cryogenic electronics is important for many mission-critical applications such as quantum computer and quantum sensing interfaces [1], space exploration electronics [2], and high-performance cryogenic servers [3]. Technology Computer-Aided Design (TCAD) offers a very cost-effective way of exploring the design space of cryogenic electronics and there has been tremendous progress in the calibration, modeling, and simulation methodologies of cryogenic electronic simulation recently [4][5]. However, cryogenic *ab-initio* quantum transport simulation, which is important to study devices with L_G < 20nm and, particularly, *their sub-threshold behavior, is still difficult and not studied systematically.*

In this paper, we propose a simulation methodology for robust and accurate *ab-initio* quantum transport simulation down to 3K of an n-type Si nanowire. This is important to understand the subthreshold swing (SS) at cryogenic temperature. We show that for L_G = 10nm, the SS is fully dominated by direct tunneling at cryogenic temperature, which *is the first time to* ***be demonstrated using ab initio simulation****, to the best of our knowledge. We propose a method to* ***achieve more than 2X speed up in the simulation and achieve convergence at high gate biases.***

Simulation Setup

QuantumATK is used in this study [6]. A Silicon unit cell is cleaved in <100> direction, repeated 2 times in both transverse directions (A and B), and passivated with hydrogen using sp^3 hybridization (Fig. 1). The structure is relaxed using LCAO with GGA for exchange-correlation and PseudoDojo for pseudopotential. The relaxed structure is repeated 60 times in the longitudinal direction (C) to form a 30nm nanowire and a cylindrical gate of L_G is added (Fig. 1). The equivalent oxide thickness (EOT) is about 1.6nm. An electrode length of 5.4nm is used for both the source and drain electrodes. The same LCAO settings are used for quantum transport simulation. Since it is a gate-all-around (GAA) structure, the Dirichlet boundary condition is applied in all directions in the Poisson solver. To achieve a reasonable simulation time, the parallel conjugate gradient is used for the Poisson solver. The n-type Source/Drain is doped to a level such that the Fermi level is above the conduction band using the compensation charge method. The channel is doped with p-type.

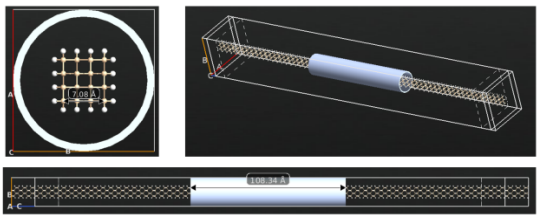


Figure 1: The cross-section and 3D views of the nanowire simulated. L_G = 10nm is shown. Gate radius is 9Å. Gate dielectric is vacuum with EOT~1.6nm.

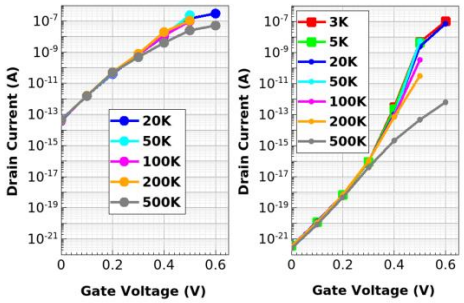


Figure 2: I_DV_G of the nanowire with L_G=5nm at T = 300K (Left) and T = 3K (Right) with various T_B. V_D=50mV.

Energy Broadening and Source/Drain Doping

Typical *ab initio* calculation uses the electron ground state energy for calculations and, thus, the quantities calculated such as the transmission coefficient and Local Density of States (LDOS) are for 0K. However, the convergence can be very bad due to the discontinuity in electron occupation at 0K. Often, energy broadening is applied to the Fermi-Dirac (FD) distribution in the calculation to smear out the electron distribution for better convergence. Note that this is *not* the FD at the electrodes. Eq. (1) shows the FD equation with T_B being the broadening temperature and kT_B representing the energy broadening. The occupation probability of a state at energy ϵ is,

$$f = \left(1 + e^{\frac{\epsilon-\mu}{kT_B}}\right)^{-1} \quad (1)$$

where k and μ are the Boltzmann constant and Fermi energy, respectively. At room temperature simulations, often T_B as large as 1000K is used for good convergence. Fig. 2 shows the I_DV_G simulated with various T_B and electrode temperatures, T. Note that T represents the operating temperatures of the device as it determines the electron distribution at the electrodes, and T_B is just the parameter for electron structure calculations. For T = 300K, consistent results are obtained with T_B < 200K. However, T_B < 5K is required for T = 3K simulation. Therefore, T_B = 5K is used in this study.

While setting T_B = 5K worsens the convergence, it brings another positive effect toward more accurate and robust simulations. It is found that a large T_B can cause Fermi level in the heavily doped source/drain to shift into the

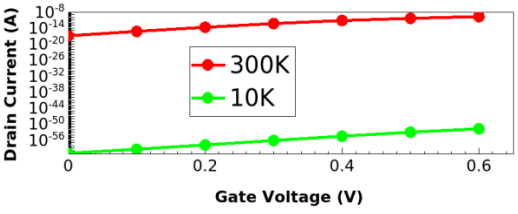


Figure 3: I_DV_G of the nanowire with L_G=5nm at T = 300K and T = 10K when the Fermi level is 80meV below the CB with S/D doping of 10^{21}cm^{-3} and T_B=1000K. V_D=50mV.

bandgap. Fig. 3 shows that when the doping is 10^{21}cm^{-3} and T_B = 1000K, the Fermi level is below the CB by 80meV and there is finite current for T = 300K but almost no current for T = 10K or below. However, using a large S/D doping will worsens the convergence due to the large doping gradient between the S/D and the channel. Fortunately, smaller T_B provides more accurate simulation and smaller S/D doping is required to keep the Fermi level above the conduction band. Fig. 4 shows the S/D doping concentration required to keep the Fermi level 18meV above the CB at various T_B. With T_B=5K, only 5×10^{20}cm^{-3} is required. Fig. 5 shows the LDOS of the device studied with V_D = 0V and V_G = 0V and the electrodes' Fermi levels.

Origin of SS Saturation in L_G=10nm Device

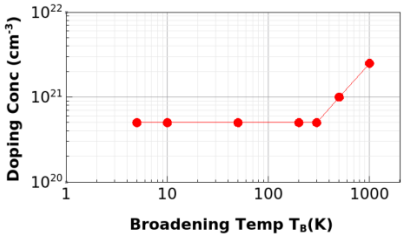


Figure 4: Source/Drain doping concentration requied to keep the Fermi level 18meV above the CB.

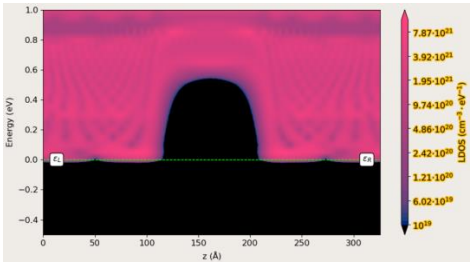


Figure 5: LDOS of the nanowire with V_G = 0V. T_B = 5K is used. L_G=10nm.

The SS slope of the GAA nanowire is then studied at various electrode/operation temperatures, T, using T_B = 5K. To further improve convergence, T_B = 300K is used as an initial solution at V_G = 0V and V_D = 50mV. Fig. 6 shows the I_DV_G curves. *The full IV curve is possible because of the robust setup.* Fig. 7 shows the speed up by using lower S/D doping which is enabled by using T_B = 5K. It can be seen that the subthreshold slope stops following the Boltzmann statistics (SS = ln(10)kT/q) when T = 77K. The subthreshold current is found to be dominated by the tunneling current at T < 77K by fitting to the tunneling equation of a rectangular barrier. There is no advantage going to T = 3K operation in terms of off-state leakage unless a very low V_{TH} device is needed as there is still a 10X reduction in leakage at 0.5V from 77K to 3K (equivalent to having V_{TH}=0.1V).

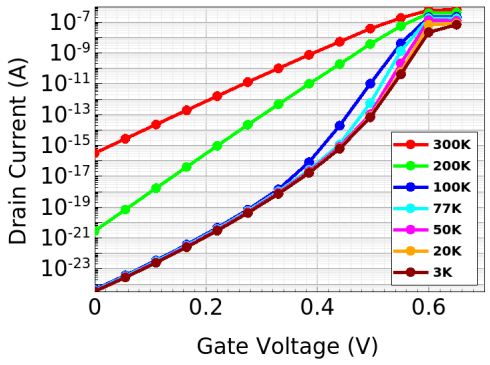


Figure 6: I_DV_G of the nanowire with L_G=10nm at various T with T_B = 5K and S/D doping of 5×10^{20}cm^{-3}.

Conclusions

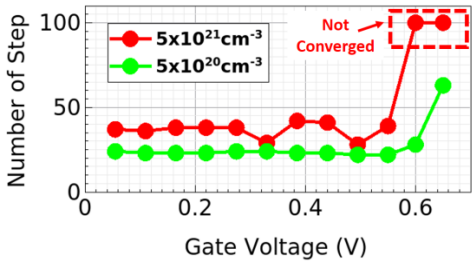


Figure 7: Number of steps to achieve convergence at each gate bias for the I_DV_G in Fig. 6.

For the first time, we completed a full I_DV_G simulation of an L_G=10nm nanowire at T=3K. We proposed a methodology to achieve an accurate and robust cryogenic *ab initio* quantum transport simulation down to 3K. It is proved that the leakage is dominated by tunneling at 3K but there is about a 10X reduction in leakage if a V_{TH} = 0.1V device is used.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 2046220. The authors thank Anders Blom from QuantumATK for his support and discussion.

References

[1] E. Charbon et al., "Cryo-CMOS for Quantum Computing," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 13.5.1-13.5.4.

[2] M. J. Gong et al., "Design Considerations for Spin Readout Amplifiers in Monolithically Integrated Semiconductor Quantum Processors," 2019 IEEE Radio Frequency Integrated Circuits Symposium (RFIC), Boston, MA, USA, 2019, pp. 111-114.

[3] R. L. Patterson et al., "Electronics for Deep Space Cryogenic Applications," Proceedings of the 5th European Workshop on Low Temperature Electronics, Grenoble, France, 2002, pp. 207-210.

[4] A. Beckers, " Cryogenic MOSFET Modeling for Large-Scale Quantum Computing," Ph.D. dissertation, Faculté des sciences et techniques de l'ingénieur, Lausanne, EPFL, 2021.

[5] H. Y. Wong, "Calibrated Si Mobility and Incomplete Ionization Models with Field Dependent Ionization Energy for Cryogenic Simulations," 2020 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), Kobe, Japan, 2020, pp. 193-196.

[6] Søren Smidstrup et al., "QuantumATK: an integrated platform of electronic and atomic-scale modelling tools," J. Phys.: Condens. Matter 32 (2020) 015901 (36pp).

A compact physical expression for the static drain current in heterojunction barrier CNTFETs

Manojkumar Annamalai and Michael Schröter, *Senior Memeber IEEE*

Chair for Electron Devices and Integrated Circuits, Technische Universität Dresden, Germany

Carbon nanotube field-effect transistors (CNTFETs) are promising candidates for future high-speed/-frequency system-on-chip applications, including highly linear radio-frequency (RF) front-ends integrated with digital CMOS electronics [1-3]. Analog high-frequency (HF) CNTFETs fabricated at wafer-scale recently achieved cut-off frequencies around 100 GHz [4] which is at the same level as that of RF CMOS. For designing analog RF CNTFET circuits, an accurate compact model is needed, which must accurately describe the drain current and tube charge over a wide bias range. Existing compact expressions for the drain current that assume ohmic contacts [5] are not suitable for describing actual fabricated CNTFETs which exhibit a heterojunction barrier (HB) at the contact-CNT interface so that the drain current is determined by tunneling through that barrier. Compact drain current formulations based on an energy independent tunneling transmission factor [6] or based on either Airy functions or even numerical integration of the Landauer equation that include the impact of HBs [7-9] are computationally expensive for circuit simulation and contain model parameters that are difficult to determine by measurements. Given these challenges, this paper presents a novel compact physics-based description for the drain current of HB CNTFETs that is based on a closed-form analytical solution of the Landauer equation. The single continuous formulation is valid for both Fermi- and Boltzmann statistics without the necessity of partitioning the energy (and hence the bias) region and includes the energy dependence of the transmission factor.

For the investigation pursued here, a single-tube within the 3D unit cell structure, which may be part of a multi-tube CNTFET, with top gate electrode located symmetrically between source and drain contact as shown in Fig. 1(a) was simulated. The simulation is facilitated by a computationally efficient simulation approach based on augmented drift-diffusion transport (aDD) with WKB boundary conditions and coupled with a 3D Poisson solver [10]. The schematic band diagram, shown in Fig. 1(b), presents the different current components of the total source injected current as well as the influence of gate and drain voltages on the band edges. The schematic also shows the possible multiple scattering of injected carriers between the barriers.

At low injection (subthreshold region) the conduction band edge underneath the gate at z_t exceeds the top of the barrier, i.e. $W_{FS} + q\Phi_{bn}$, resulting in the thermionic emission of carriers. The Landauer integral is solved analytically for the thermionic current $I_{n,thS}$ by assuming an energy and bias independent transmission factor. $I_{n,thS}$ increases with gate-source (GS) voltage until the flat band case, i.e. the band edge equaling $W_{FS} + q\Phi_{bn}$. From thereon $I_{n,thS}$ becomes bias independent and a function of just the barrier height. With further increase in GS voltage, the conduction band edge drops below $W_{FS} + q\Phi_{bn}$ and tunneling through the HB ($I_{n,tuS}$) dominates the total source injected current, and the transmission factor T_{nS} in the Landauer integral becomes strongly energy dependent. Since the shape of the HB, rather than its extent, determines the transmission of the carriers, it is convenient to approximate the band edge near the channel end with an exponential function, which agrees well with the simulation results as shown in Fig. 2(a), and use a single energy point at the middle of the conduction band edge W_{Ct} coupled to the gate potential. The Landauer integrand, given by the function $f_{tu} = f_n T_{nS}$, with f_n as the Fermi-function, is displayed on the right of Fig. 2(a). Calculating f_{tu} using the source Fermi level W_{FS} and the analytical expression of T_{nS} based on an exponential barrier profile agrees well with the result from device simulation. Realizing that the energy dependence of the transmission distribution has the shape of the Gaussian, it can be approximated by $f_{tu} \approx f_{tu,p}\exp(-((W-W_{p,tu})/\sigma_{tu})^2)$; the transmission distribution function reaches its peak value $f_{tu,p}$ at $W_{p,tu}$ with σ_{tu} as its standard deviation. Since the transmission distribution and its Gaussian approximation decrease to negligible values towards the boundaries of the integration interval, the integration limits can be replaced by $\pm\infty$, yielding after multiplication with $4q/h$ the closed-form solution for the source injected tunneling current $(4q/h)\sqrt{\pi}\sigma_{tu}f_{tu,p}$. The integral is pictorially represented by the shaded region in Fig. 2(a).

Based on the observed behavior of $W_{p,tu}$ and σ_{tu} characterizing the Gaussian approximation of the f_{tu}, shown in Fig. 2(b), their bias dependence, especially the smooth transition from the subthreshold to the high-injection region, is modeled accordingly. The bias dependent function $f_{tu,p}$ requires the evaluation of both the Fermi function and the transmission factor at the peak energy $W_{p,tu}$ of the transmission distribution. The drain injected current component is calculated similarly, and the total current formulation includes multiple-backscattering at the hetero-barriers. The effect of carriers scattering along the channel at higher drain-source (DS) voltages is taken into account by a computationally simple function [11]. The resulting compact physical current expression shows good agreement with the device simulation results as shown in Fig. 3(a-d). Furthermore, excellent agreement of drain current and transconductance with measured data from the multi-tube HF CNTFET in [12] is observed in Fig. 3(e, f), confirming the suitability of the new current formulation also for fabricated devices. The developed formulation has been integrated into our compact CNTFET model [11].

[1] M. Hartmann *et al*., *IEEE J. of Microwaves*, vol. 1, no. 1, pp. 275–287, 2021.
[2] Y. Cao *et al*., *Topics in Current Chemistry*, vol. 375, no. 5, p. 75, 2017.
[3] M. Schröter *et al*., *IEEE J. of the Electron Devices Society*, vol. 1, no. 1, pp. 9–20, 2013.
[4] C. Rutherglen *et al*., *Nature Electronics*, vol. 2, no. 11, pp. 530–539, 2019.
[5] R. Marani, G. Gelao and A.G. Perri, *Microelectronics J.*, vol. 44, no. 1, pp. 33–38, 2013.

[6] Y. Zhang *et al.*, *IEEE Trans. on Electron Devices*, vol. 68, no. 12, pp. 6571–6579, 2021.
[7] R.A. Vega, *IEEE Trans. on Electron Devices*, vol. 53, no. 4, pp. 866–874, 2006.
[8] P. Michetti and G. Iannaccone, *IEEE Trans. on Electron Devices*, vol. 57, no. 7, pp. 1616–1625, 2010.
[9] I. Bejenari, M. Schröter and M. Claus, *IEEE Trans. on Electron Devices*, vol. 64, no. 9, pp. 3904–3911, 2017.
[10] S. Mothes and M. Schröter, *IEEE Trans. on Nanotechnology*, vol. 17, no. 6, pp. 1282–1287, 2018.
[11] M. Schröter *et al.*, *IEEE Trans. on Electron Devices*, vol. 62, no. 1, pp. 52–60, 2015.
[12] Y. Cao *et al.*, *ACS Nano*, vol. 10, no. 7, pp. 6782–6790, 2016.

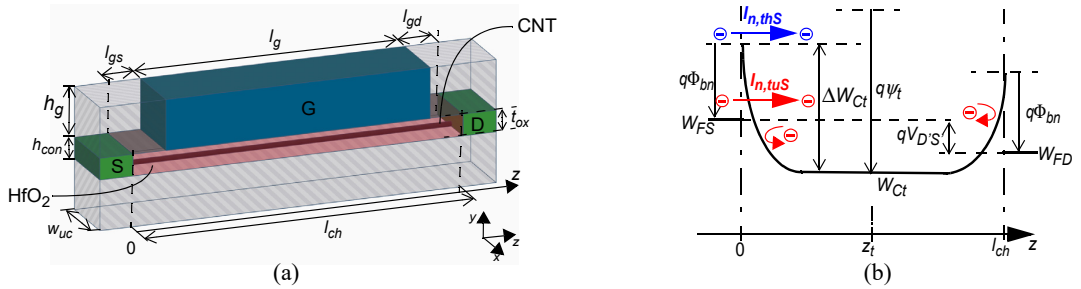


Fig. 1: (a) 3D view of the simulated top-gate CNTFET unit cell. (b) Schematic band diagram at low $V_{D'S'}$ and high $V_{G'S'}$ with $q\Phi_{bn}$ as barrier height

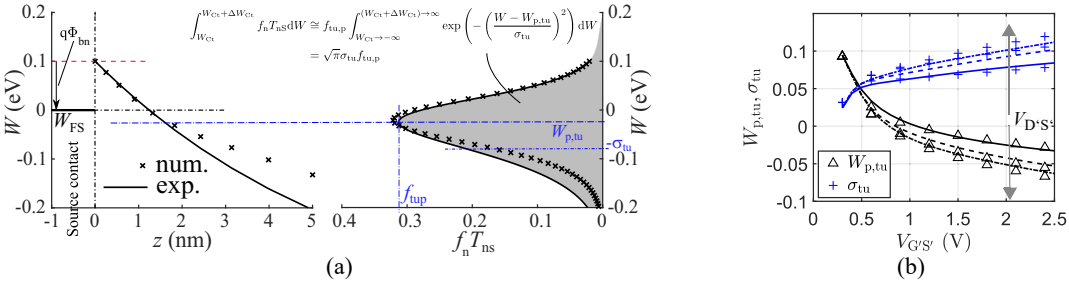


Fig. 2: (a) Spatial dependence of conduction band edge near the source end of the channel (left) and transmission distribution function versus energy (right). (b) Bias dependence of the variables determining the Gaussian approximation determined from f_{tu} with numerical T_{nS} (symbols) and analytical T_{nS} (lines).

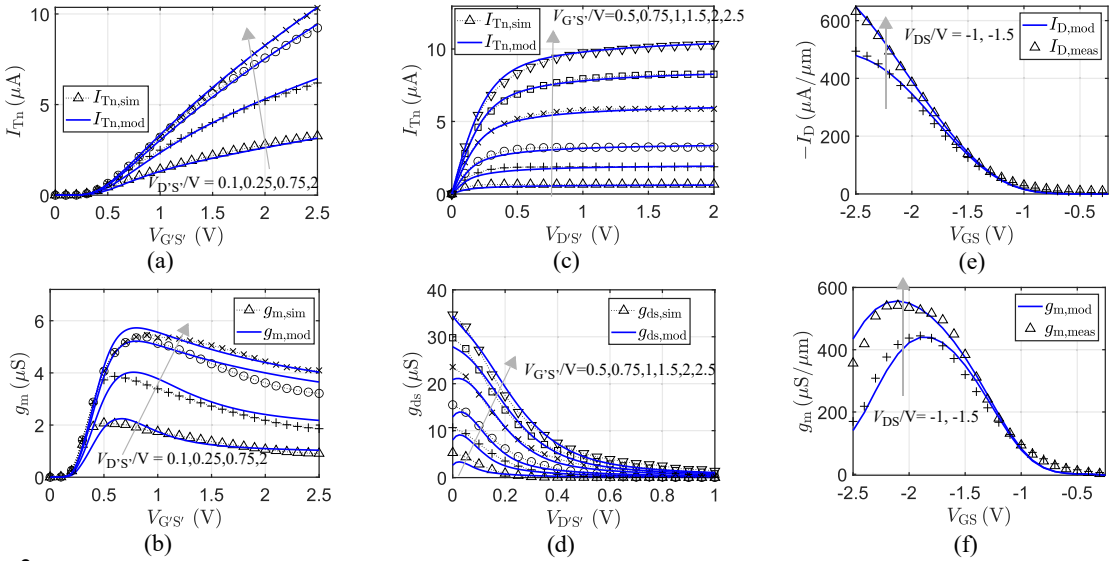


Fig. 3: Comparison between compact model (lines) and numerical simulation (symbols) for (a), (b) transfer characteristic and transconductance; (c), (d) output characteristics and conductance as a function of internal GS and DS voltage. Comparison between compact model (lines) and measurements of a HF CNTFET (symbols) for (e) drain current and (f) transconductance as a function of external GS and DS voltage.

Enabling medium thick gate oxide devices in 22FDX® technology for switch and high-performance amplifier application

Tom Herrmann[1], Alban Zaka[1], Zhixing Zhao[1], Binit Syamal[1], Wafa Arfaoui[1], Ruchil Jain[1], Ming-Cheng Chang[1], Sameer Jain[2], Shih Ni Ong[3]
[1]GlobalFoundries Dresden [2]GlobalFoundries Malta [3]GlobalFoundries Singapore

1) Introduction

22FDX® platform constitutes an excellent choice for RF and mmWave application combined with superior low power logic performance and non-volatile memory [1]. Extension of the platform offering entails trade-offs between device performance and application-specific reliability criteria. In this work, we focus on the optimization of the medium-thick gate oxide devices from a device performance perspective, targeting high-performance amplifier and switch applications. We make use of the well-known gate length scaling effect combined with a modified Source/Drain doping profile [2,3].

2) TCAD setup and calibration

Following the calibration strategy of the thin gate oxide devices [4], medium thick devices have been calibrated based on DC, AC and RF data in SProcess and SDevice including all relevant process steps. **Fig. 1** shows the architecture and doping profile of the NFET device featuring an undoped channel and nwell below the BOX. **Fig.2** shows the TCAD to HW correlation for under-, regular- and over-drive device (different V_{DD} and gate length – L_G) and well type below the BOX (SLVT-nwell / LVT-pwell) for typical DC parameters on a single finger structure. The standard quantum-corrected drift-diffusion with a unique process and device model parameter set is used throughout the study [5]. **Fig.3** finally shows TCAD vs. HW comparison in terms of F_t, F_{max} and gm for the under-drive device with a satisfactory agreement on a multi-finger structure. With confidence in the initial TCAD ability to mimic measurements, it was used to optimize tradeoffs and obtain state-of-the-art performance.

3) Scaling effects in 22FDX® platform

To ensure low HCI degradation medium thick oxide devices have graded junctions to reduce the electrical field at the drain to gate edge. **Fig.4** shows the doping profile of two processes, P-1 and P-2 with a long, respectively short, extension of the high-doping region towards the gate. The graded junction shows up to 30% peak electric field reduction under worse HCI conditions. On the other hand, having a more abrupt doping profile improves the threshold voltage roll-off and reduces DIBL, thus enabling shorter gates with good electrostatic behavior (**Fig. 5**). Since the under-drive devices have margin to pass the HCI criteria (driven by the lower V_{DD}) the more abrupt doping profile of P-2 combined with a shrink in the gate length will still be sufficient. Furthermore, the Drain leakage current and the device linear resistance at constant drain bias (R_{on}) is significantly reduced (**Fig. 6**). The R_{on} reduction is driven by a better channel mobility due to the absence of dopants in the portion of channel under the gate. P-1 and P-2 process minimize R_{on} with gate length reduction having same slope, which means P-1 performance is worse, since V_{tlin} scaling is stronger. Looking at the RF FoMs in **Fig.7**, F_t shows a stable upward trend with device scaling and a negligible sensitivity to the gate overlap. F_{max} shows some gain as we scale the device, particularly with P-2. However, F_{max} is saturating for the shortest gates. Here the 2D TCAD simulations are using a gate length dependent lumped resistance at the gate contact to mimic the vertical and horizontal resistance of the 3D structure. **Fig.8** shows that significant improvements of F_t and F_{max} are obtained at high drain current and less, but still visible, at low drain currents for a combination of shorter gates and P-2. The minimum gate length possible will additionally be defined by the application specific reliability criterion.

Fig. 9 depicts the C_{off}*R_{on} FoM for switch application [6], which is improved with shorter gates and P-2 process. Furthermore, the leakage current is only marginally increased at the same threshold voltage and the drain breakdown voltage can be maintained at a high level visible in the C_{off}*R_{on}/V_{max} metric. Using LVT device construction would further reduce the leakage current at marginally lower C_{off}*R_{on} and C_{off}*R_{on}/V_{max}, later FoM saturating for P-2 at about -30nm gate length. The simulation data shows encouraging results for a switch device candidate and the device implementation in switch design will be further studied.

Despite the RF FoMs previously discussed, R_{on}, gate capacitance (C_{GG}) and leakage current are also important parameters which amplifier designers pay close attention to. The steeper doping profile of P-2 is reducing the gate-drain capacitance (C_{GD}) and the gate scaling benefits the overall C_{GG}, illustrated in **Fig. 10**. R_{on} and Idoff have already been shown to reduce with gate length scaling and P-2 process. Combining the back bias (BB) capability of FDSOI devices, further reduction of the drain leakage (reverse BB) at a little cost of Ron can be achieved, makes it an attractive feature for amplifiers (**Fig. 10**).

References:

[1] M.Wiatr, et al: “22FDX™ Technology and Add-on-Functionalities”, 2019, ESSDERC
[2] T.Nicoletti, at al: “The Impact of Gate Length Scaling on UTBOX FDSOI Devices: the Digital/Analog Performance of Extension-less Structures”, 2012, ULIS

[3] A. Majumdar, et al: "Gate Length and Performance Scaling of Undoped-Body Extremely Thin SOI MOSFETs", 2009, EDL
[4] T.Herrmann, et al: "RF performance improvement on 22FDX® platform and beyond", 2019, SISPAD
[5] E.M.Bazizi, et al.: "Versatile technology modeling for 22FDX® platform development", 2017, SISPAD
[6] T.McKay, et al.: "Advances in Silicon-On-Insulator Cellular Antenna Switch Technology", 2009, SIRF

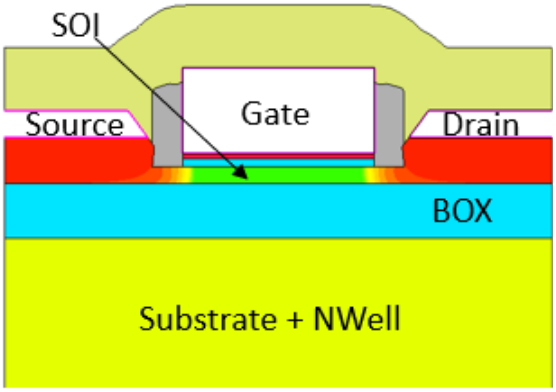


Fig.1: TCAD device structure and doping profile

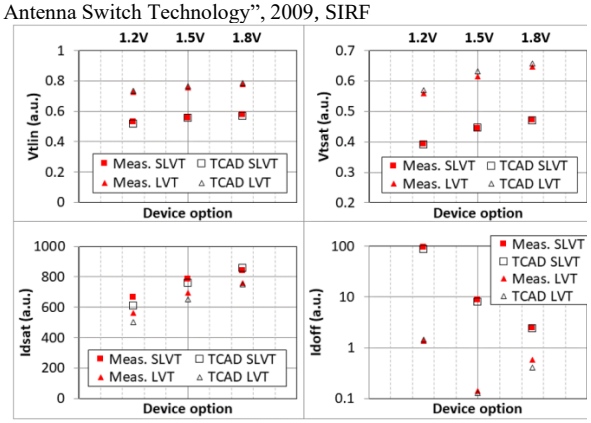
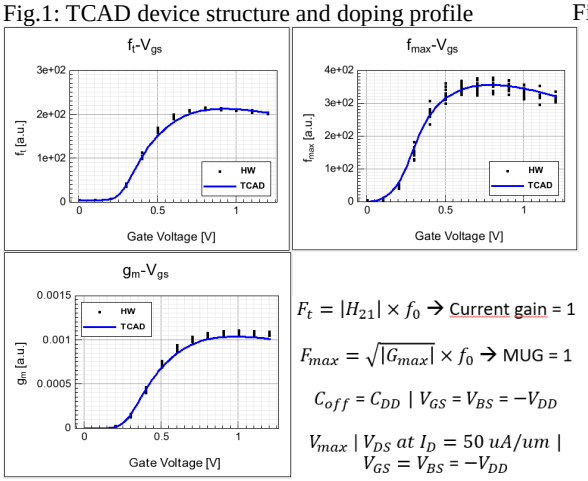


Fig.2: DC calibration (TCAD – open / HW – filled sym.)



$F_t = |H_{21}| \times f_0$ → Current gain = 1

$F_{max} = \sqrt{|G_{max}|} \times f_0$ → MUG = 1

$C_{off} = C_{DD}$ | $V_{GS} = V_{BS} = -V_{DD}$

V_{max} | V_{DS} at $I_D = 50\ uA/um$ | $V_{GS} = V_{BS} = -V_{DD}$

Fig.3: RF calibration and FoM extraction methods (TCAD - solid line / HW - squares)

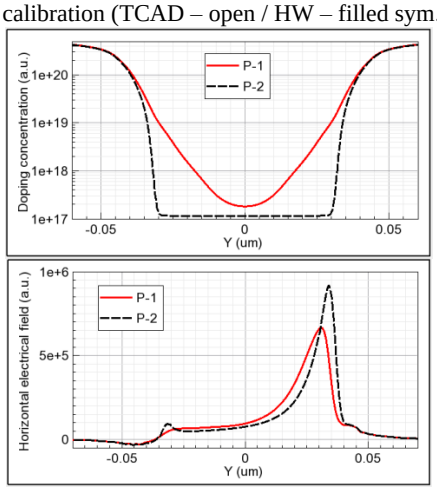


Fig.4: Doping profile and horizontal electrical field at V_{DS}=V_{GS}=1.2V for P-1 and P-2 process

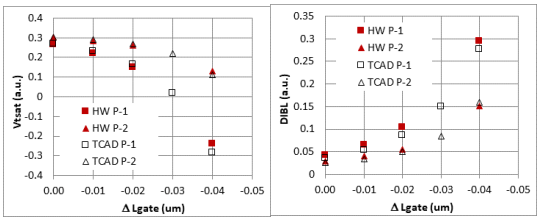


Fig.5: Saturation threshold voltage and DIBL

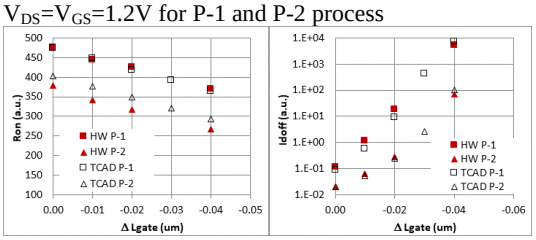


Fig.6: R_{on} and Drain leakage current

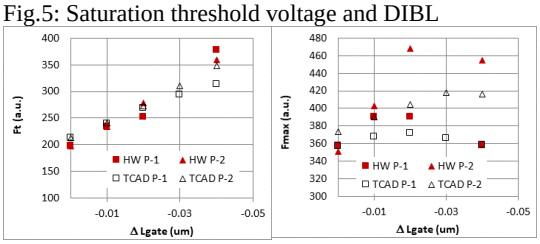


Fig.7: F_t and F_{max} scaling

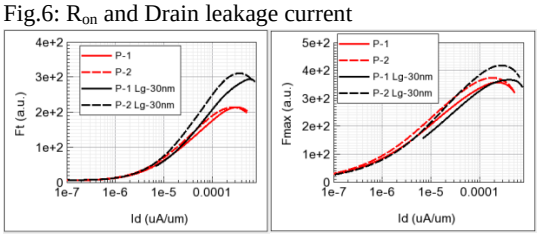


Fig.8: F_t and /F_{max} vs. drain current for P-1/P-2 process and two gate length at V_{DS}=1.2V

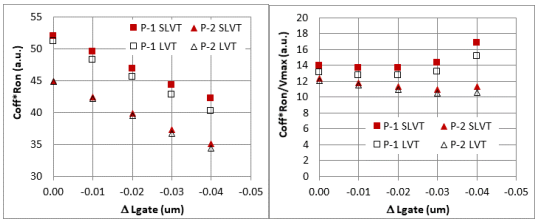


Fig.9: C_{off}*R_{on} and C_{off}*R_{on}/V_{max} for P-1/P-2 process and n/pwell back gate doping

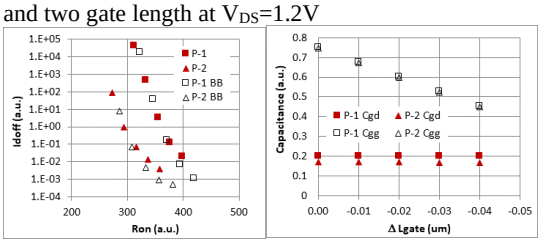


Fig.10: I_{doff}/R_{on} for V_{BS}=0V/-0.8V(BB) for different gate length and C_{GD} (V_{GS}=0V)/C_{GG} (V_{GS}=1.2V)

Non-Quasi-Static Modeling and Methodology in Fully Depleted SOI MOSFET for L-UTSOI model

S. Martinie[1], O. Rozeau[1], HyoEun Park[2], Sungjoon Park[2], P. Scheer[3], S.El Ghouli[3], A. Juge[3], H. Lee[2], T. Poiroux[1]
1) Univ. Grenoble Alpes, CEA, LETI, 38000 Grenoble, France; 2) Samsung, South Korea; 3) STMicroelectronics, France.

Abstract: *The Fully-Depleted Silicon On Insulator (FDSOI) technologies are deployed for a wide range of RF applications at high frequency operation. Usually, the surface potential based compact models of MOSFET (implemented in SPICE) are built within the Quasi-Static (QS) approximation, which means that continuity equation has no temporal dependence and consequently carriers are not dynamically affected along the channel. This paper describes the recent improvements of L-UTSOI standard model by the adaptation of the NQS-RTA model from literature to enhance it. This new feature has been validated using the segmented model approach and silicon experimental data.*

Keywords: ***NQS, L-UTSOI, compact model, SPICE, FDSOI.***

1) NQS SPICE modeling: between complexity and efficiency.

Under a high-frequency operation, the charge modulation in MOSFET may no longer follow the signal instantaneously, leading to a so-called Non-Quasi-Static (NQS) effect. It has been largely explored in literature for SPICE, analytical or small-signal modeling where different models or methodologies have been proposed [1-8].

From a SPICE model usage point of view, the most accurate solution is to break down the MOSFET in N equal channel segments using an appropriate methodology for parameter setting and physical effects activation [7-8]. Independently of the runtime that is, strictly speaking dependent of the number of transistors (ie. segments) solved in SPICE netlist. The main limitation of the segmented model is the parameter extraction adapted to all physical phenomena (such as SCE, DIBL, GIDL, saturation velocity …) that cannot be topologically segmented (such as low field mobility). For example, constant part of source or drain resistance is externally set (RSW1 parameter in L-UTSOI [10] is replace by external resistance at source and drain of equivalent circuit), but the corresponding voltage dependencies are more difficult to handle (RSGO parameter in L-UTSOI [10]).

Another method is the spline collocation-based approach [4-5] where the 1D current continuity equation is solved. In practice, the user chooses the number of channel partitions that defines the number of collocation points. The channel charge density is then approximated by a cubic spline through these collocation points. Each collocation point (up to 9 in PSP model through SWNQS parameter [5]) corresponds to an additional RC branch where an equivalent number of internal nodes (in addition to modeling equation) is necessary to achieve a good accuracy with respect to the segmented model. Even if the parameter extraction is not affected (only extra mobility parameter MUNQS is used for the transition frequency modulation), the spline collocation-based approach is difficult to apply to models such as L-UTSOI without compromising the runtime efficiency and the robustness of the simulations.

Finally, the most popular simplified approach is the use of Relaxation Time Approximation (RTA), which reduces the extra runtime and improves convergence, but with lower accuracy [1-3]. The principle is to use an extra sub-circuit to filter the charge, and incorporate a time dependence. Initial implementation [1-2] used a QS calculation of the partition from the inversion charge to drain charge, but as explained in [3], it is more efficient to apply the RTA to both the inversion and drain charges.

2) Segmented model vs RTA NQS in L-UTSOI: model and discussion.

L-UTSOI is the first available standard compact model able to describe the behavior of an ultra-thin body and BOX FDSOI transistor including strongly inverted back interface [10-11]. Regarding NQS modeling, we firstly illustrate the segmented model case with ten segments (cf. fig 1.a) by using similar methodology as [7]. As depicted in figure 2, we reproduce exactly the TCAD simulation results on the DC (fig. 2.a) and frequency dependent figures of merit (fig. 2.b & 2.c). However, as explained in the previous part, this approach has some limitations (runtime, complexity...). Thus, we propose to use the NQS-RTA in L-UTSOI. The global topology is illustrated in figure 1.b where the RTA equations (1) is solved for Q_I and Q_D by the circuit simulator thanks to the use of internal nodes. The relaxation time is given by equation (2), similar to [2-3]. An empirical pole is added to improve a little bit the accuracy through the partition parameter $\alpha \in [0,1]$ with an associated relaxation time K_α. Thus, equations (3-4) represent the temporal part of the terminal currents.

Figure 2.b & 2.c illustrate clearly the good accuracy of NQS-RTA model up to twice the transition frequency. Interestingly on figure 2.d & 2.e, the model is also able to capture the back bias dependence, including the back channel inversion charge since Q_i naturally includes this effect. Figure 3 is a comparison on long and wide experimental device on 28nm FDSOI. We show here for a full range of V_{DS} & V_{GS}, the capability of the NQS-RTA model to reproduce measurements. Note that QS extraction (low frequency) is not illustrated here, but follows the usual extraction flow, only KDRIFT and KDIFF are extracted.

Figure 4 shows the simulation of "killer" NOR gate circuit which is a good benchmark of NQS interest [4]. The large peak, observed in QS condition due to a non physical current from the top transistor, is strongly reduced when we use the NQS model. As expected, this peak modulation is directly dependent on the model's parameters setting. In practice, since NQS implementation is an "end of code" charge filtering, which is relatively "light" in term of implementation (simulation of 11 stages Ring Oscillator Fan-Out 4 shows a runtime increase around 20%). In conclusion, the adaptation of NQS-RTA model in L-UTSOI is the right compromise between runtime, accuracy and simplicity in comparison to the segmented model, while preserving the description of back interface inversion.

References: [1] M. Chan et, vol 45, no. 4, april 1998; [2] D. Navarro et al, TED, vol 53, no. 9, September 2006; [3] Z. Zhu et al, TED, vol 59, no. 5, May 2012; [4] H. Wang et al, TED, vol 53, no. 9, September 2006; [5] PSP 103.7 Verilog-A and documentation. http://www.cea.fr/cea-tech/leti/pspsupport; [6] A-S. Porret et al; TED, vol ED-32, no. 11, Nov 1985; [7] A.J. Sholten et al, IEDM 1999; [8] M. Bucher et al, SSE 52 (2008); [9] M. Bagheri & Y. Tsividis, TED, vol 62, no. 9, September 2015 [10] T. Poiroux et al, TED, vol 62, no. 9, September 2015; [11] L-UTSOI 103.5 Verilog-A and documentation https://www.cea.fr/cea-tech/leti/l-utsoisupport.

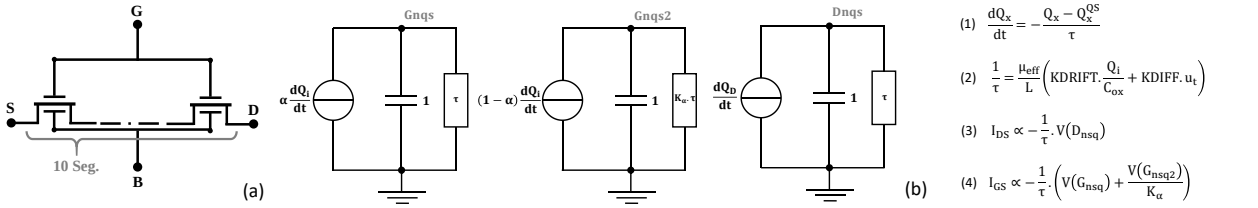


Fig. 1: (a) segmented topology and (b) equivalent circuit used to emulate NQS-RTA model and superimposed to the MOSFET. Corresponding notations for equations: Q_x charge at node x, QS index is for Quasi-Static, μ_{eff} effective mobility, Q_i inversion charge, u_t thermal voltage, C_{ox} oxide capacitance & L channel length.

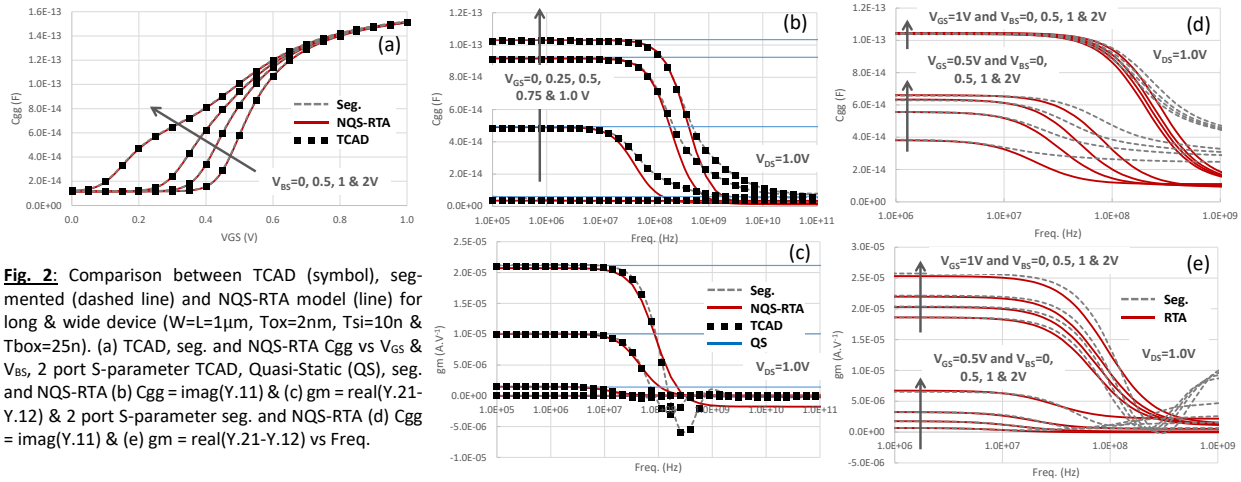


Fig. 2: Comparison between TCAD (symbol), segmented (dashed line) and NQS-RTA model (line) for long & wide device (W=L=1µm, Tox=2nm, Tsi=10n & Tbox=25n). (a) TCAD, seg. and NQS-RTA Cgg vs V_{GS} & V_{BS}, 2 port S-parameter TCAD, Quasi-Static (QS), seg. and NQS-RTA (b) Cgg = imag(Y.11) & (c) gm = real(Y.21-Y.12) & 2 port S-parameter seg. and NQS-RTA (d) Cgg = imag(Y.11) & (e) gm = real(Y.21-Y.12) vs Freq.

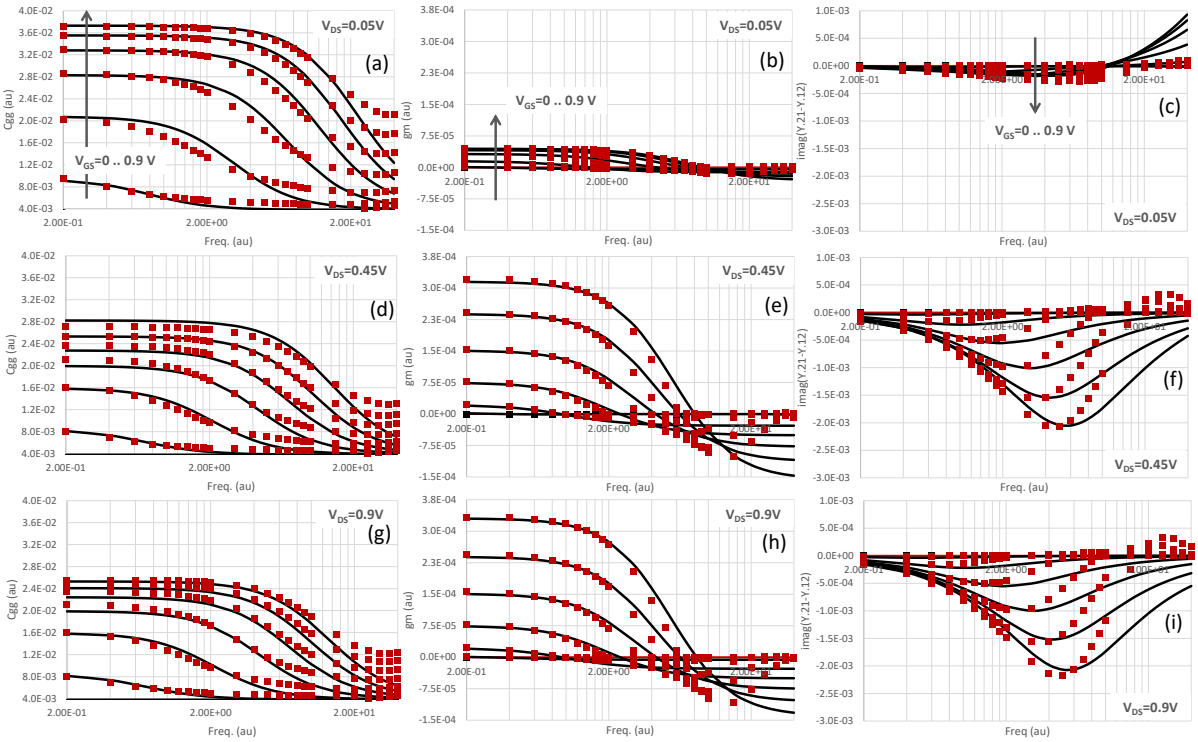


Fig. 3: Experiments from 28nm technology (symbol) vs NQS-RTA model (line) for long & wide device on 2 port S-parameter measurement and simulation: Gm = real(Y.21-Y.12), Cgg = imag(Y.11) and imag((Y.21-Y.12) vs Frequency for V_{GS}=0.0 .. 0.9V at V_{DS}=0.05V (a, b & c), V_{DS}=0.45V (d, e & f) and V_{DS}=0.9V (g, h & i).

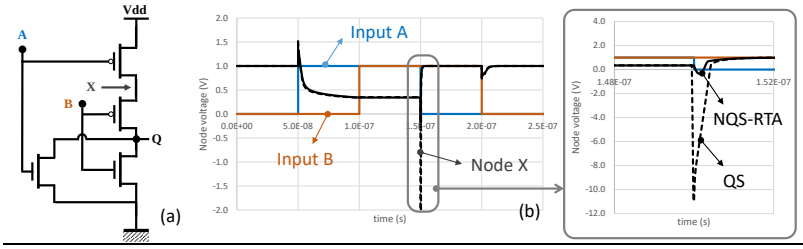


Fig. 4: Practical illustration of NQS-RTA model on circuit simulation. (a) "Killer" NOR Gate circuit used to illustrate the interest of NQS-RTA model and (b) comparison of the simulated voltage at node X of the killer NOR gate using the Quasi-Static (QS without any frequency dependence) and the NQS-RTA models.

String-level Compact Modeling Based on Channel Electrostatic Potential for Dynamic Operation of 3D Charge Trapping Flash Memories

Sunghwan Cho
Department of Semiconductorl and Display Engineering
Sungkyunkwan University
Suwon 16419, Republic of Korea
joboss9999@gmail.com

Byoungdeog Choi
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon 16419, Republic of Korea
bdchoi@skku.edu

***Abstract*—In this paper, we have developed a channel-potential-based circuit model of charge trapping flash (CTF) memory devices with 3D vertical gate-all-around (GAA) structure. To achieve accurate performance and dynamic operations, BSIM-CMG Verilog-A model using cylindrical GAA structure and equivalent circuit describing single modeling unit composed of metal gate, charge trapping layer, tunneling/blocking oxide, and poly-Si layer are used. For determination of the accurate potential level of boosted channel and dynamic operation, SPICE simulation results show good agreement with experimental data of incremental step pulse programming (ISPP) and technology computer-aided-design (TCAD) simulation results of positively boosted channel electrostatic potential.**

Keywords—Compact model, BSIM-CMG, 3D CTF memory

I. Introduction

For the scaling down of the memory devices, charge trapping flash (CTF) memory with 3D vertical structure is regarded as a good alternative, solving the capacitive-coupling issues and the reliability with a larger storage area compared to planar short channel structure [1][2]. And the growing demands of mobile computing and data centers continue to drive the need for high capacity and performance of 3D vertical NAND flash technology. However, scaling down of flash memory cell size still has a limitation in improving the performance of devices as the cell-to-cell interference is increased [3]. Furthermore, CTF memory with 3D vertical structure shows much more complex behavior in channel operation than the conventional planar memory devices due to the poly-Si channel with floated body. This brings an essential requirement of accurate compact model for dynamic circuit-level simulation to design optimized and advanced multilevel memory devices. Although compact model based on circuit-level operation of Flash memory has been reported previously [4][5], it is mainly focused on conventional planar structures and several models proposed recently are developed for the numerical simulation or calculation in order to get the insights of the characteristics [6][7][8].

In this work, we have developed an accurate compact model for 3D CTF memory devices based on channel electrostatic potential with gate-all-around (GAA) structure. For predicting accurate level of channel electrostatic potential, measurement of incremental step pulse programming (ISPP) and technology computer-aided-design (TCAD) simulation are used. And extended netlist for describing the equivalent circuit of cylindrical GAA layers composed of metal gate, charge trapping layer, tunneling oxide, blocking oxide and poly-Si channel gate-all-around (GAA) structure is used for enhanced model behavior and dynamic circuit-level operation.

II. Simulation and Results

A. Device Structure

Fig. 1 shows the schematics of modeling structure which is investigated in this work. Based on definition of single modeling unit in gate-all-around structure as shown in Fig. 1(a), corresponding equivalent circuit has been proposed describing cylindrical MOS transistor with stacked capacitors of tunneling oxide, charge trapping nitride layer and blocking oxide in series to the gate of conventional BSIM-CMG model [9], as shown in Fig. 1(b). In addition, current sources are attached for describing leakages of FN tunneling through the tunneling/blocking oxide and band to band tunneling between adjacent cells.

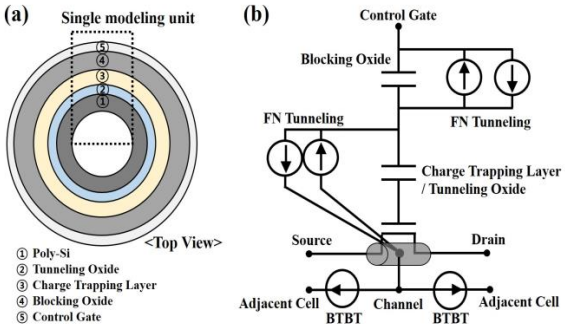


Fig. 1. (a) Description of single modeling unit and (b) schematic of equivalent circuit in the proposed model of 3D CTF memory devices.

B. Simulation Results and Discussions

Fig. 2 and 3 show the TCAD simulation results of electrostatic potential of poly-Si channel at inhibited string during program operation, comparing with SPICE results. Based on well-predicted leakage model and cylindrical capacitance calculation, proposed model has good agreements with TCAD simulation results, as shown in Fig. 3. In addition, optimized result of potential lowering during program execution caused by BTBT leakage makes it available for circuit designer to determine the optimized bias conditions of adjacent cells in regard to disturbance. BTBT leakage mechanism, which dominantly occurs at narrow depletion region and in the case of a large potential difference between victim and neighbor cell, is used as below

$$I_{BTBT} = \frac{\alpha * \Delta V_{CH}}{\left(\frac{\beta * W_{DEP}}{\Delta V_{CH}}\right)^{1.5} * \exp\left(\frac{\beta * W_{DEP}}{\Delta V_{CH}}\right)}$$

where W_{DEP} is the depletion width, ΔV_{CH} is the potential difference of victim cell comparing upper (or lower) channel, and α and β are optimized parameters.

The comparison between the proposed model and the measured threshold voltage during incremental step pulse programming (ISPP) operation is shown in Fig. 4, indicating that the boosted potential also can be defined as the threshold voltage difference between programmed and inhibited behavior. With good agreement to experimental data, it is confirmed that accurate compact model is developed to predict and analyze the characteristics of 3D charge trapping flash memory during dynamic operation.

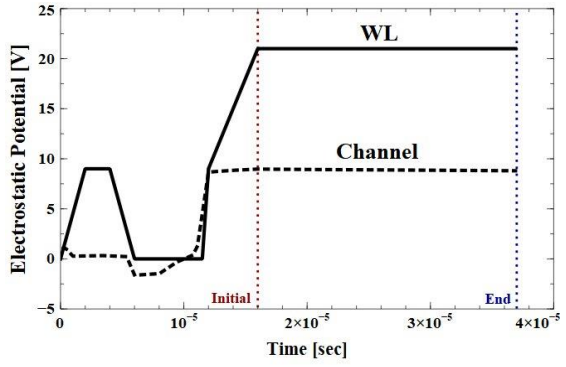


Fig. 2. Positively boosted potential of poly-Si channel simulated by SPICE during program operation at inhibited string.

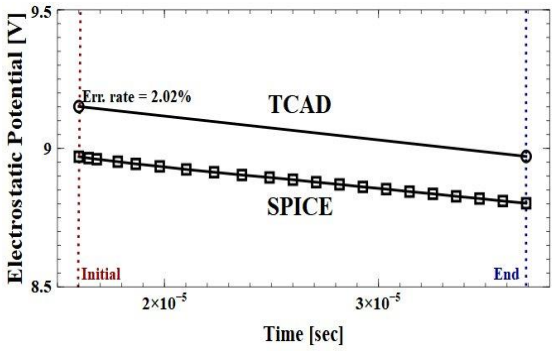


Fig. 3. SPICE results of boosted potential showing good agreement with TCAD simulation results at initial point and between initial and end point which indicates electrostatic potential lowering by BTBT leakage during program execution period.

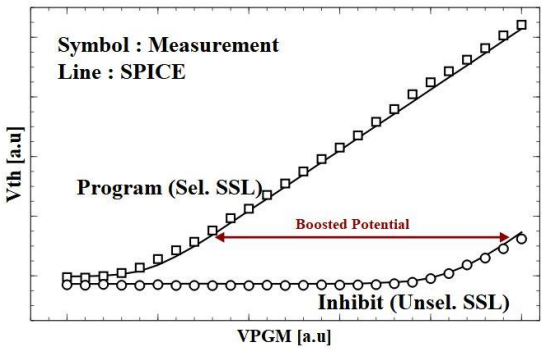


Fig. 4. SPICE simulation results during ISPP operation in the programmed and inhibited string showing good prediction for characteristics of 3D CTF memory devices.

III. Conclusion

In this paper, we developed a compact model based on channel electrostatic potential for the 3D charge trapping flash memory devices using extended BSIM-CMG model and verified the model through the result of measurement and TCAD simulation. To predict the electrostatic potential of poly-Si channel accurately, optimized parameters are extracted by good agreement with measured behavior and TCAD simulation. Good predictions of ISPP operation using proposed model make it available to optimize design and analyze failures of 3D CTF memory devices with a faster simulation method.

References

[1] J. M. Koo and T. E. Yoon, "Vertical structure NAND Flash array integration with paired FinFET multi-bit scheme for high-density NAND Flash application.", IEEE 2008 Symposium on VLSI Technology Digest of Technical Papers, 120-121 (2008).

[2] Akira Goda, "3D NAND technology achievements and future scaling perspectives", IEEE Transactions on Electron Devices, vol. 67, no.4, April 2020.

[3] K. T. Park and M. G. Kang, "A zeroing cell-to-cell interference page architecture with temporary LSB strong and parallel MSB program scheme for MLC NAND flash memeroies.", IEEE J. Solid-State Circuit vol. 43, no.4, 919-928 (2008).

[4] A. Torsi, "A Program disturb model and channel leakage current study for sub-20nm NAND flash cells.", IEEE Transactions on Electron Devices, vol. 58, no.1, 11-16 (2011).

[5] J. W. Jeon, I. H. Park, "Accurate compact modeling for sub-20nm NAND flash cell array simulation using PSP model.", IEEE Transactions on Electron Devices vol. 59, no.12, 3503-3509 (2012).

[6] W. C. Chen and H.T. Lue, "A Physics-based Quasi-2D Model to understand the wordline (WL) interference effects of junction-free structure of 3D NAND and experimental study in a 3D NAND Flash test chip", IEDM (2018)

[7] Gerardo Malavena and Aurelio Mannara, "Compact modeling of GIDL-assited erase in 3D NAND Flash string", Journal of Computational Electronics, vol. 18, 561-568 (2019)

[8] Gerardo Malavena and Aurelio Mannara, "Investigation and compact modeling of the time dynamics of the GIDL-assited increase of the string potential in 3D NAND Flash array", IEEE Transactions on Electron Devices, vol. 65, no. 7, July (2018)

[9] [Online] Available: http://bsim.berkeley.edu

Semiconductor workforce development through immersive simulations on nanoHUB.org

Gerhard Klimeck, Tanya Faltens, Daniel Mejia, Alejandro Strachan, Lynn Zentner, Michael Zentner*,
Network for Computational Nanotechnology, Purdue University
*San Diego Supercomputing Center, UCSD

Over 160,000 nanoHUB users have run over 7 million simulations in Apps mostly focused on semiconductor devices and materials modeling. nanoHUB created nano-Apps before Apple created Apps for the iPhone and made scientific codes usable for a much larger user group. Most scientific tools strive to be comprehensive in solving "any" simulation problem in a specific problem range. That comprehensiveness limits the use to experts, who require extensive training. nanoHUB has instead focused on delivering a spectrum of Apps (over 700 now) that individually have a limited capability focused on a PN-junction, MOSFET, or nanowire while the underlying tool could of course solve a much wider set of problems. We assembled some of these Apps that are essential for specific courses into small sets such as ABACUS (crystals, bandstructure, drift-diffusion, pn-junctions, BJTs, MOScaps, MOSFETs) [1]. The usability results are stunning. Our user analytics prove that over half of the simulation users participate in structured education through homework/project assignments. We can identify classroom sizes and detailed tool usage [2,3]. We can begin to build mind-maps of design explorations and assess depth of explorations for individuals and classes. While parts of academia struggled to innovate curricula, we have measured the median first-time App insertion into a class to be less than six months. Over 180 institutions have utilized nanoHUB in their curriculum innovation in over 3,600 classes. 2 million nanoHUB visitors explore lectures and tutorials annually. With such a community presence we believe nanoHUB is the platform of choice to deliver online modeling, simulation, virtual environments, and lectures for the US initiative on workforce development [4]. This presentation will overview some of the nanoHUB impact metrics and turn towards a tool recitation session. In the recitation a brief overview of ABACUS will be given and the audience may request other tool demonstrations or exploration.

[1] https://nanohub.org/groups/abacus ABACUS - Assembly of Basic Applications for Coordinated Understanding of Semiconductors. A one-stop-shop for teaching and learning semiconductor fundamentals.
[2] Krishna Madhavan, Michael Zentner, Gerhard Klimeck, **"Learning and research in the cloud"**, Nature Nanotechnology 8, 786–789 (2013)
[3] TEDx Talk, Klimeck, "Mythbusting Scientific Knowledge Transfer with nanoHUB.org", https://www.youtube.com/watch?v=PK2GztIfJY4 .
[4] https://nanohub.org/groups/semiconductoreducation Semiconductor workforce development homepage on https://nanoHUB.org .

Dr. Gerhard Klimeck is a Professor of Electrical and Computer Engineering at Purdue University; Director of the Network for Computational Nanotechnology; Reilly Director of the Center for Predictive Materials and Devices. He helped to create nanoHUB.org, the largest virtual nanotechnology user facility serving over 2.0 million global users, annually. Dr. Klimeck is a fellow of the Institute of Physics (IOP), the American Physical Society (APS), the Institute of Electrical and Electronics Engineers (IEEE), the American Association for the Advancement of Science (AAAS), and the German Humboldt Foundation. He has published over 525 printed scientific articles; he has been recognized for his co-invention of a single-atom transistor, quantum mechanical modeling theory, and simulation tools. His NEMO5 software has been used since 2015 at Intel to design nano-scaled design transistors. The nanoHUB team was recently recognized by a top 100 by R&D award - Making simulation and data pervasive.

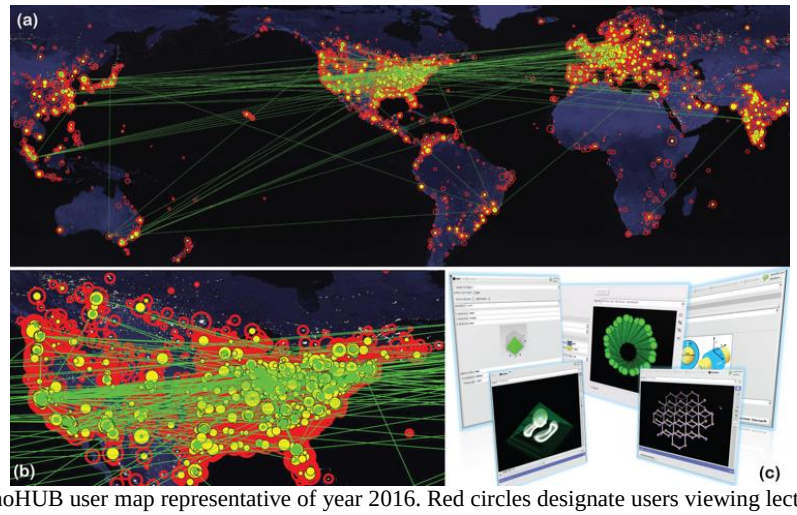


Figure 1 (a) nanoHUB user map representative of year 2016. Red circles designate users viewing lectures, tutorials, or homework assignments. Yellow dots are simulation users. Green dots indicate 4,100+ authors of 1,700+ scientific publications citing nanoHUB through 2016. Dot sizes correspond to number of users, and lines show author-to-author connections proving research collaboration networks. (b) U.S. enlarged. (c) a collage of typical nanoHUB interactive tool sessions and 3D-rendered interactively explorable results (quantum dots, carbon nanotubes, nanowires).

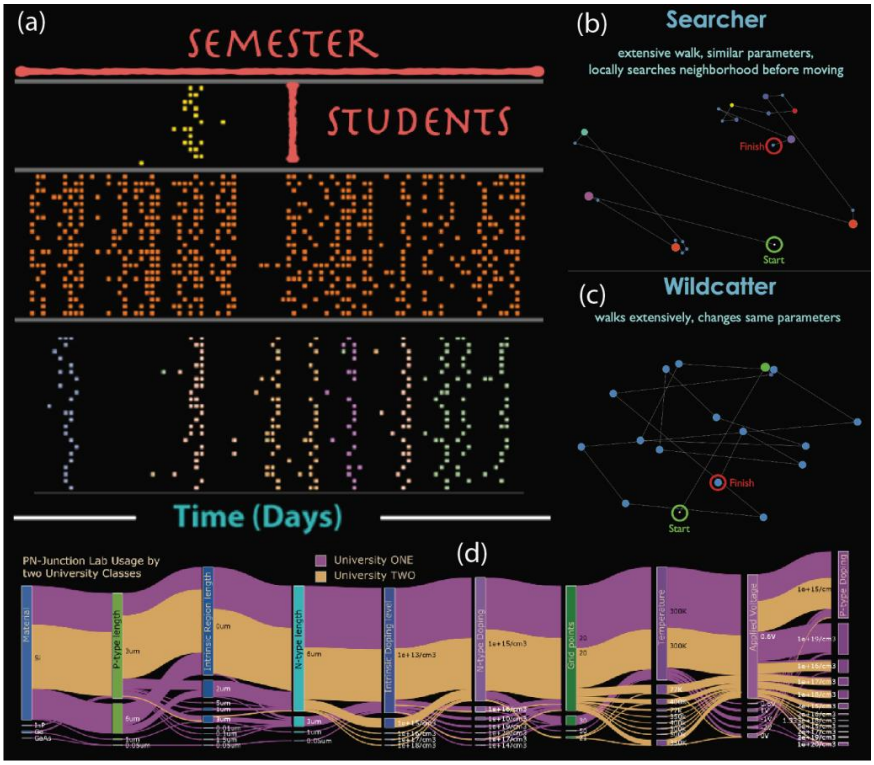


Figure 2 (a) nanoHUB user behaviors over time identified as classes. (b,c) Specific behavior of 2 *PN Junction Lab* users classified as Searcher and Wildcatter. (d) Collective behavior of 2 classes using *PN Junction Lab*.

1D Drift-Diffusion transport in 2D-material based FETs with vertical contacts

A. Toral-Lopez *,[1] E.G. Marin,[1] F. Pasadas,[1] M.C. Pardo,[1] J. Cuesta,[1] F.G. Ruiz,[1] and A. Godoy[1]
[1] *PEARL Laboratory, Dpto. de Electrónica y Tecnología de Computadores, Universidad de Granada (Spain).(*atoral@ugr.es)*

Abstract

A simulation scheme for 2D-material (2DM) based FETs with vertical (top) contacts is presented. The approach is able to capture the physics of 2DM-contact system at the macroscopic level while facilitating the integration of the material and interface properties coming from lower levels of abstraction, and setting a first step towards a multi-scale modelling of the 2DM-contact. The model relies on two physical parameters: the vertical conductance in the metal/semiconductor heterojunction, and the metal work function. The role of each of these parameters on the device behaviour is analysed, assessing their impact in the contact resistance.

I. Introduction

After almost two decades of research in 2DMs, electrical contacts still remain as a technological challenge to be controlled, and thus hindering the exploitation of the expected potential of 2DM-based devices [1]. The analysis of contacts implemented as vertical heterostructures, where the current injection is perpendicular to the carrier transport in the channel [2], has been extensively faced from *ab-initio* techniques [3] or ballistic approaches [4, 5] while the numerical semi-classical device-level modelling is less explored. Here, we present a Drift-Diffusion (DD) transport scheme able to capture the physics of 2DM-contact system at the macroscopic level.

II. Numerical Methods

The numerical simulation of the device comprises the self-consistent solution of the electrostatics, using the Poisson equation, combined with the Continuity and DD transport equations. The DD equation is implemented under the Fermi-level-gradient formulation for electrons ($J_\mathrm{n} = n\mu_\mathrm{n}\nabla E_{\mathrm{F,n}}$) and holes ($J_\mathrm{p} = p\mu_\mathrm{p}\nabla E_{\mathrm{F,p}}$). For the continuity equation, the current density is assumed to have only an in-plane component in the 2DM except in those regions were contacts are defined, where an out-of-plane injection, and therefore a current is enabled, as indicated in Figure 1.

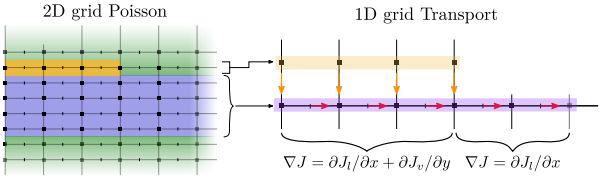


FIG. 1. Modelling of the vertical contacts in the structure. The subscript l (v) stands for longitudinal (vertical).

Each contact is characterised by its bias (energetic position of the Fermi level); the work function of the metal ϕ_m, which impacts on the boundary condition for the electrostatic potential in the region where the contact is defined ($V_\mathrm{BC} = V_\mathrm{bias} - (\phi_\mathrm{m} - \chi_\mathrm{s})/q$); and the vertical mobility μ_v that models the quality of the contact. χ_s corresponds to the electron affinity of the 2DM.

After the self-consistent simulation of the device for a specific bias range, the macroscopic resistance associated the contact is evaluated employing the charge density profile in the semiconductor located under the contact, which is self-consistently evaluated with the electrostatic potential, as well as the in-plane and out-of-plane mobility in this region. Figure 2 summarizes the approach.

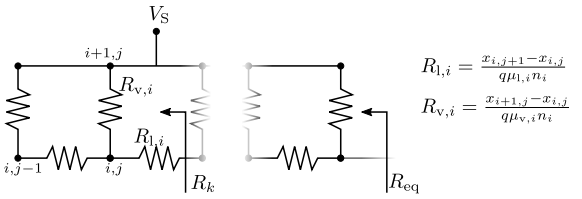


FIG. 2. Calculation of the contact resistance for the source contact. The nodes of the simulation grid are connected by a resistor, resulting in a set of resistors in a cascade configuration that provide the equivalent resistor observed from the edge of the channel, i.e. the contact resistance.

At the microscopic level, the contact behaviour can be better understood in terms of the transference length ($l_{\mathrm{v,S/D}}$), i.e. the length of the region where almost all the vertical current injection takes place which typically ranges a few nanometers ([6]). Here, this length is defined by the position at which the magnitude of the vertical component of the current density decays a 99%, as depicted in Figure 3. The reference value to evaluate this decay is the one at the edge of the contact, which corresponds to the maximum of the vertically injected current density.

III. Results

In order to evaluate the proposed simulation scheme, we consider the structure depicted in Figure 3, that corresponds to a single gate MoS_2 MOSFET, with source and drain defined by two 10nm-long contacts.

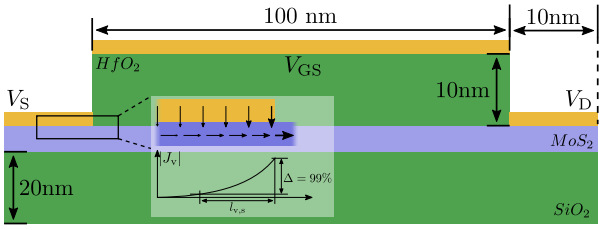


FIG. 3. Structure of the MoS_2 based MOSFET simulated to evaluate the impact of the vertical contact mobility and its relation with the contact resistance.

The gate bias V_GS is set to 0.25V, so that the device is in the on state, while three different V_DS values are considered: 0.1V, 0.3V and 0.5V. These bias conditions are employed along with different values of the out-of-plane mobility μ_v and metal work function $\phi_\mathrm{m} - \chi_\mathrm{s}$.

Figure 4 shows the output current I_DS as a function of μ_v for all the scenarios evaluated evidencing that a poorer contact quality, i.e. a lower μ_v, results in a lower I_DS, as

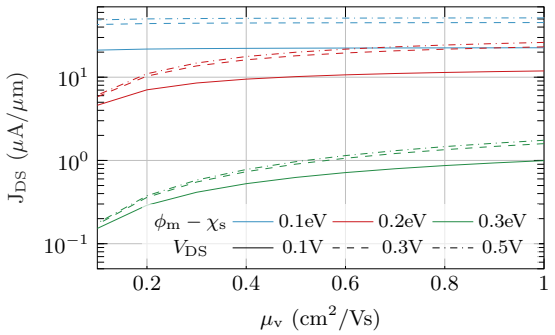


FIG. 4. $I_{DS} - \mu_v$ profile for different values of ϕ_m (line colors) and V_{DS} (line style).

could be expected. Notably, ϕ_m shows a higher impact on I_{DS}: when the barrier height ($\phi_m - \chi_s$) is increased, the output current is significantly decreased by one or two orders of magnitude. In order to evaluate how the quality of the contact, i.e. mobility and barrier height, impact the injection length, Figure 5 shows the $l_v - \mu_v$ profile for the source (Figure 5a) and drain (Figure 5b) contacts.

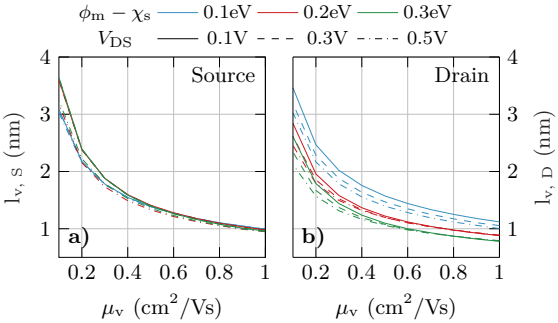


FIG. 5. $l_v - \mu_v$ profile for the source (a) and drain (b) contacts obtained using different ϕ_m and V_{DS} values.

The source injection length is mainly impacted by μ_v, keeping always in the range of a few nms, rather than by the barrier height, with the better the contact quality (high μ_v) the smaller the region where charge is injected. At the drain, $l_{v,D}$, shows slight change with ϕ_m and V_{DS} that could be related to the electrostatic interaction between the gate and drain contact that modifies the transport and charge concentrations conditions near the drain edge.

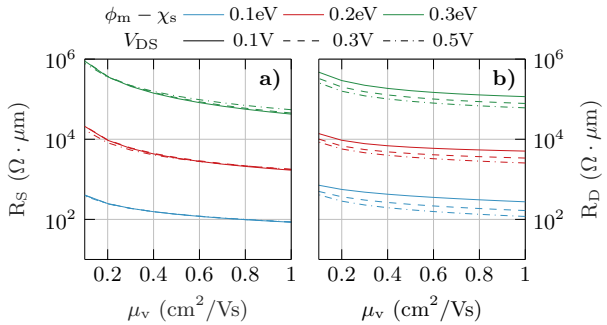


FIG. 6. Source (a) and drain (b) contact resistances as a function of the vertical mobility for different values of $\phi_m - \chi_s$ and V_{DS}.

As for the macroscopic description, Figure 6 depicts the behaviour of the contact resistance as a function of μ_v. In this case, both source and drain are significantly modified when changing $\phi_m - \chi_s$ and μ_v. R_D also changes to a large extent when drain bias is modified. To verify this dependence we evaluated the longitudinal charge density profile under source and drain contacts for different V_{DS} values for the $(\mu_v, \phi_m) = (1\ \text{cm}^2/\text{Vs}, \chi_s + 0.1\text{eV})$ scenario. The data depicted in Figure 7 show that the carrier density under the drain contact significantly varies with V_{DS}, which is not the case of the source contact, explaining the asymmetric behaviour of both contact resistances.

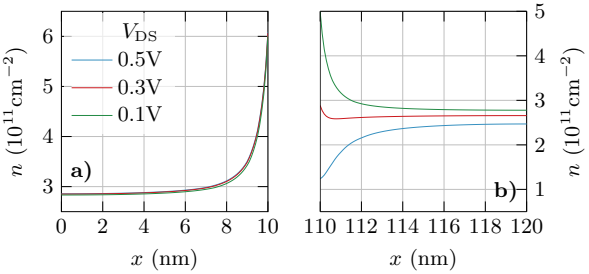


FIG. 7. Longitudinal charge density distribution under source (a) and drain (b) contacts for different V_{DS} values and $(\mu_v, \phi_m) = (1\ \text{cm}^2/\text{Vs}, \chi_s + 0.1\text{eV})$.

IV. Conclusions

We have introduced a self-consistent platform for the simulation of vertical (top) contacts in 2DMs based FETs. The impact of the parameters employed to model the contacts (μ_v and $\phi_m - \chi_s$) on the output current, injection length and contact resistances have been analysed, observing that I_{DS} is mainly modified by $\phi_m - \chi_s$ while l_v is mostly influenced by μ_v. Notably both parameters have a significant impact on the contact resistance. In addition to that, we observe that the contact resistance at the drain, R_D, also depicts a noticeable dependence with V_{DS}. This behaviour is linked with the strong modulation of the carrier density under the drain contact when its bias is modified.

Acknowledgments

This work was supported by MCIN/ AEI/ 10.13039/ 501100011033 through the project PID2020-116518GB-I00, by FEDER/ Junta de Andalucía Consejería de Transformación Económica, Industria, Conocimiento y Universidades through the Projects B-RNM-375-UGR18, A-TIC-646-UGR20 and PY20-00633, and by European Union through the Project WASP under Contract 825213. A.Toral-Lopez acknowledges Plan Propio programme from UGR, and J.Cuesta acknowledges Spanish Ministry of Universities (FPU2019/05132).

[1] Y. Guo et al. *Nano Research*, 14(12):4894–4900, 2021. URL http://dx.doi.org/10.1007/s12274-021-3670-y.
[2] A. Allain et al. *Nature Materials*, 14:1195–1205, 2015. ISSN 1476-1122. URL http://dx.doi.org/10.1038/nmat4452.
[3] J. Kang et al. *Physical Review X*, 4:031005, 2014. ISSN 2160-3308. URL http://dx.doi.org/10.1103/PhysRevX.4.031005.
[4] D. Marian et al. *Physical Review Applied*, 8(5), 2017. URL http://dx.doi.org/10.1103/physrevapplied.8.054047.
[5] E. G. Marin et al. *ACS Nano*, 14(2):1982–1989, 2020. URL http://dx.doi.org/10.1021/acsnano.9b08489.
[6] D. S. Schneider et al. In *2021 Device Research Conference (DRC)*. IEEE, 2021. URL http://dx.doi.org/10.1109/drc52342.2021.9467156.

A generalizable, uncertainty-aware neural network potential for GeSbTe with Monte Carlo dropout

Sung-Ho Lee*, Valerio Olevano†, Benoît Sklénard*‡
*Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France
†Université Grenoble Alpes, F-38000 Grenoble, France
CNRS, Institut Néel, F-38042 Grenoble, France
‡E-mail: benoit.sklenard@cea.fr

***Abstract*—A Bayesian neural network potential (NNP) achieved with the Monte Carlo dropout approximation method is developed for GeSbTe alloys. The Bayesian NNP is shown to be more generalizable than its classical counterpart, yielding reasonable predictions on structures that are not directly in the training configurations, and is able to output uncertainty estimates for the predictions.**

I. Introduction

GeSbTe (GST) alloys are interesting materials for a variety of technological applications like phase change memories or photonic devices [1]. Atomistic simulations are often used to survey their properties and to optimize their composition, but studying finite temperature properties such as phase transition or thermal conductivity often requires long simulations of extended systems that are too expensive for *ab initio* methods. Neural network potentials (NNP), trained on Density Functional Theory (DFT) references, can be used to overcome this limitation, which were recently shown to enable simulations at scales that were previously impossible at a near-DFT level accuracy [2].

NNPs, however, can silently fail on structures that lie outside the learned configuration space which poses some issues when relying upon them for atomistic simulations. One solution is to apply the Bayesian paradigm to the neural network to capture the model's uncertainty along with its predictions. This has previously been demonstrated for the different phases of carbon [3], but never to more complex materials containing more than one chemical species.

In this work, a Bayesian NNP using the Monte Carlo (MC) dropout technique will be used on GST materials to evaluate its ability to estimate the predictive uncertainty as well as its superior generalizability compared to classical NNPs.

II. Methodology

A. Classical and Bayesian HDNNP

The High Dimensional NNP (HDNNP) [2] is a type of NNP that is composed of a set of Atomic Neural Networks (ANN) that each takes one atom of a given system as input and outputs a so-called "atomic energy", whose sum provides the total energy of the system.

The ANNs are fully-connected layers that, classically, contain point weights and output point estimates. In contrast, Bayesian neural networks place a probability prior over the weights and the estimates are obtained as a distribution by sampling from the posterior. In practice, for complex functions like neural networks, computing and sampling from the posterior is intractable. MC dropout [4] is a simple and efficient method to approximately sample from the posterior by using dropout at inference time.

Dropout was originally developed as a regularization technique in which a fraction of nodes of a neural network are randomly "dropped" (ie. set to zero) during training to prevent overfitting [5]. In MC dropout, this random dropping is also applied at inference time, and multiple stochastic forward passes through the network is performed, probabilistically selecting different nodes at each iteration to produce a distribution of outputs. The prediction and the associated uncertainty is then obtained as the mean and standard deviation of this distribution [4].

B. Computational Details

To generate the datasets, *ab initio* molecular dynamics (MD) calculations based on Density Functional Theory (DFT) were carried out using the VASP code [6], [7] at temperatures from 300 K to 1500 K at 300 K intervals for 20 ps for hexagonal GeTe, Sb_2Te_3 and $Ge_2Sb_2Te_5$. 200 snapshots were then selected for each trajectory to compute accurate energies and forces.

Our in-house library FFLearn was used for the NNPs employing the symmetry functions proposed by Behler [2] to encode the atomic environments of each atom for neural network input. All NNPs were defined to have 2 hidden layers with 15 nodes each. For the Bayesian NNPs, nodes were dropped with a probability of 10%, and sampled 100 times at inference time.

III. Results and Discussion

For a neural network potential to be practicable for atomistic simulations it must have the ability to generalize as much as possible to previously unseen configurations, and, more importantly, to recognise when it cannot.

Since a neural network potential can be thought of as a complex function modelling the potential energy surface (PES), regions on the PES that are outside the learned regions can lead to erroneous predictions. Being able to recognise when this occurs would make the NNP more reliable.

A. Comparison of Generalizability

In order to assess the generalizability, the two NNPs (Bayesian and classical) were trained on the hexagonal GeTe 300 K, 600 K, and 900 K datasets (the training set) and made to predict the energies of the 1200 K and 1500 K snapshots

Table I. RMSE (meV/atom) of hexagonal GeTe between the DFT energy and the NNP predictions

	Validation	1200 K	1500 K
Bayesian	3.18	53.1	1339.5
Classical	1.49	561.5	9185.0

Table II. RMSE (meV/atom) of $Ge_2Sb_2Te_5$ between the DFT energy and the NNP predictions at all temperatures

Temperature (K)	300	600	900	1200	1500
Bayesian	70.4	50.0	32.0	14.6	25.6
Classical	992.7	964.2	933.6	902.2	455.5

(the test set). Root mean squared error (RMSE) between the predictions and the reference DFT energies was used as the accuracy metric, and the results are summarised in Table I.

The higher temperatures of the test sets was used to emulate structures drawn from outside the configuration space immediately spanned by the training set. The Bayesian NNP shows some promising results for the 1200 K dataset, with an RMSE of 53.1 meV/atom. The error becomes much greater for the 1500 K dataset which is expected to contain structures that are more amorphous-like, thus containing too many unknown configurations. Meanwhile, the classical NNP is unable to produce any valid results for either datasets.

B. Uncertainty estimation

The 1500 K dataset example in the previous section highlights the importance of being able to estimate the prediction uncertainty. Fig. 1 shows that this is achievable with the Bayesian NNP, exhibiting a clear correlation between the uncertainty estimate and the prediction error (computed as the absolute difference between the DFT reference energy and the NNP prediction). It is important to note, however, that this uncertainty is an estimate rather than a true reflection of the error, and care must be taken when interpreting it. Namely, they do not necessarily have a strictly linear relationship.

Making use of the uncertainty estimates typically involves setting a threshold, above which the structure is considered unknown to the model. Following this, an NNP-MD can, for example, raise a warning or be stopped. The identified structure can also be used to improve the potential by means of active learning or on-the-fly learning.

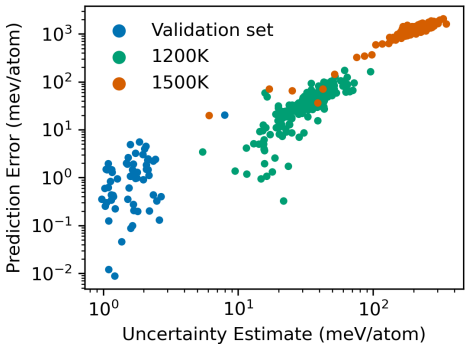


Fig. 1. Uncertainty estimate for the hexagonal GeTe validation and test sets against the prediction errors.

C. Predictions on $Ge_2Sb_2Te_5$

We now focus on $Ge_2Sb_2Te_5$ (GST225), a material that lies on the GeTe—Sb_2Te_3 tie-line on the ternary diagram. To compare the NNPs' generalizability to ternary alloys, a classical and a Bayesian NNP were each trained on the two extrema (GeTe and Sb_2Te_3) and tested on $Ge_2Sb_2Te_5$ at different temperatures. The results are summarised in Table II, and the lowest error case (1200 K) is represented graphically

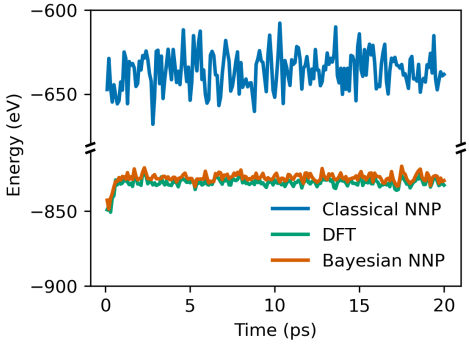


Fig. 2. Energy predictions on $Ge_2Sb_2Te_5$ at 1200 K given by the classical NNP and Bayesian NNP trained on GeTe and Sb_2Te_3, compared to the DFT reference.

in Fig. 2. Similarly to the results in Sec. III-A, the classical HDNNP fails to give meaningful predictions.

On the other hand, the relative success of the Bayesian NNP is surprising, given that it has not been supplied with any direct knowledge of the Ge–Sb or the Ge–Sb–Te relationship (or any permutation thereof). Nonetheless, these results hint at the possibility of modelling other GST compositions along the GeTe—Sb_2Te_3 tie-line with only a minimal dataset on the GST itself. As the dataset generation is the most expensive step to training NNPs, this could prove to be a very powerful development. However, more work is required to confirm these findings on different stoichiometries of GST. Furthermore, an improvement of the prediction errors, especially for the lower temperature structures, would be required before physically relevant properties can be calculated accurately.

IV. Conclusion

In this work, a Bayesian neural network potential for GeSbTe was developed using Monte Carlo dropout to approximate the Bayesian posterior. It was shown to have an improved generalizability than its non-Bayesian counterpart on hexagonal GeTe, with insightful estimates on the uncertainties of the predictions. Furthermore, it surprisingly seemed to be able to reasonably calculate the energies of hexagonal $Ge_2Sb_2Te_5$, even in the absence of information on the two body Ge–Sb and the three-body Ge–Sb–Te relationships in the training dataset. Future work will involve investigating this in more detail to better understand the mechanism behind this and to try to take advantage of this for a cost-efficient training of a general GST potential.

Acknowledgment

This work was performed using HPC resources from GENCI–IDRIS (Grant 2021-A0110911995).

References

[1] S. R. Elliott, "Chalcogenide phase-change materials: Past and future," *Int. J. Appl. Glass Sci.*, vol. 6, no. 1, pp. 15–18, 2015.
[2] J. Behler, *Chem. Rev.*, vol. 121, no. 16, pp. 10 037–10 072, Aug. 2021.
[3] M. Wen and E. B. Tadmor, *Npj Comput. Mater.*, vol. 6, no. 1, p. 124, Dec. 2020.
[4] Y. Gal and Z. Ghahramani, in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 20–22 Jun 2016, pp. 1050–1059.
[5] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.
[6] G. Kresse and J. Furthmüller, *Phys. Rev. B*, vol. 54, no. 16, p. 11169–11186, Oct 1996.
[7] G. Kresse and D. Joubert, *Phys. Rev. B*, vol. 59, no. 3, p. 1758–1775, Jan 1999.

A Machine-learning-based Multi-Objective Optimization of Stacked Nanosheet Transistors for sub-3nm technology node

Haoqing Xu[1,2], Weizhuo Gan[1,2], Lei Cao[1,2], Huaxiang Yin[1,2], and Zhenhua Wu[1,2]
[1]Institute of Microelectronics, Chinese Academy of Sciences, 100029 Beijing, China, email: wuzhenhua@ime.ac.cn
[2]School of Integrated Circuits, University of Chinese Academy of Sciences, 100049 Beijing, China

***Abstract*—Ultra-scaled transistors at the state-of-art technology node have complex nonlinear dependence of performance on its channel geometry and source/drain doping profile, which brings obstacles in the advanced technology path-finding and optimization. In this work, a machine learning-based multi-objective optimization (MOO) workflow is proposed to optimize the sub-3-nm node gate-all-around (GAA) three-layer-stacked Nanosheet transistors (NSFETs) accounting for the key performance knob of channel geometry and source/drain doping profile. The artificial neural network (ANN) is trained to model the current-voltage (IV) relationship of NSFETs from 3D Technology Computer-Aided Design (TCAD) simulation results. Based on the ANN model, MOO between threshold swing, on-off ratio, and on-state current of NSFETs is performed with adaptive weighted sum (AWS) theory. The proposed workflow efficiently suggests an optimized design window of channel geometry and doping profile of NSFETs. The proposed devices satisfy the 2025 International Roadmap for Devices and Systems (IRDS) target in terms of basic electrical characteristics.**

Keywords—Nanosheet, Machine Learning, Multi-Objective Optimization.

I. Introduction

As the down-scaling of complementary metal oxide semiconductor (CMOS) to sub-3 nm technology nodes, the GAA devices, including nanowire FETs and NSFETs, are introduced to boost gate control capability and proposed to be superior device structures for further scaling[1]. The quantum physics induced effects are inevitable in the channel and give rise to complex nonlinear dependence of performance on the transistor parameter knobs. Several device optimization options are reported in previous studies, among which the channel geometry, novel channel material, and stress are intensively studied. However, limited specifications for sub-3-nm technology node NSFETs were reported in previous studies whereas detailed geometry and doping profile optimization schemes have not been investigated thoroughly for device optimization [2]. In this work, a compact IV model of three-layer-stacked NSFETs is built by the ANN that is trained by TCAD samples. Device transfer characteristics including threshold swing SS, on-off ratio I_{on}/I_{off}, and on-state current I_{on} are extracted from an ANN-based compact model and optimized with AWS theory, from which design windows are proposed for both p-type and n-type NSFETs.

II. Experiments and Discussion

The optimization workflow is composed of three parts: TCAD simulation, machine learning, and MOO (see Fig. 1). The following are the details of the three parts.

A. TCAD Simulation

3D three-layer-stacked NSFET model is built using commercial GTS Framework TCAD software which is shown in Fig. 2(a) [3]. The doping profile is determined by the thickness of the overlap junction between source/drain L_{ov} and the variance σ (see Fig. 2(b)). The device model is calibrated to experimental results of 3nm technology node NSFET using channel geometry and S/D doping parameters as listed in Table I referring to [2] (see Fig. 3). To generate drain current I_d under different bias conditions for various NSFETs, the DOE is performed with different geometric and doping parameters as is listed in Table I. Gate voltage V_g is swept from 0.2 to -0.8V for PMOS (-0.2 to +0.8V for NMOS) and V_{DD} are set to be -0.65V for PMOS (+0.65V for NMOS). All simulation tasks are performed and the data set consists of design parameters and the I_d is cleaned for training ANN in the next steps.

B. Machine Learning

An ANN is designed for compact IV modeling as shown in Fig. 4(a). Before training, data set preprocessing is performed using Boolean for feature "MOSType" and all features and labels are normalized using standard scaling for better prediction. The data set is split 60%, 20%, and 20% for train, validation, and test sets respectively. The ANN is trained using backpropagation in the Pytorch framework [4] and its performance is shown in Fig. 4(b, c) and Fig. 5, which shows a compact IV model is well-built with high accuracy (overall R^2=0.9909).

C. Multi-Objective Optimization

SS, I_{on}/I_{off}, and I_{on} are extracted from the ANN-based compact model, and MOO among them with the AWS method (see Ref. [5]) is performed as follows. By coupling SS and on-off ratio first and coupling SS and Ion later, the minimum value of utility functions of the two combinations are calculated to find the Pareto front in objective space (Fig.6 (a)(b) for PMOS, Fig. 7(a)(b) for NMOS) and Pareto optimal solutions are obtained, whose distribution probability is shown in Fig.6(c) and Fig.7(c), from which the corresponding value of parameters at max probability is preferred for NSFET design (see Table II) and the performance of the proposed devices are validated by TCAD. The optimal NSFETs outperform the devices in 2025 and 2028 IRDS targets in terms of SS, DIBL, on-off ratio, and normalized Ion (see Table III).

III. Conclusion

In this work, a machine learning-based MOO workflow for channel geometry and doping profile of three-layer-stacked NSFETs is demonstrated. The adaptive weighted sum theory is implemented in MOO targeting device performance characteristics including SS, on-off ratio, and on-state current. An optimized set of parameters is proposed and the corresponding NSFETs performance meets the 2025 and 2028 IRDS targets.

This work was supported in part by the International Partnership Program of Chinese Academy of Sciences under Grant E1YH01X, and the NSFC under Grant 91964202.

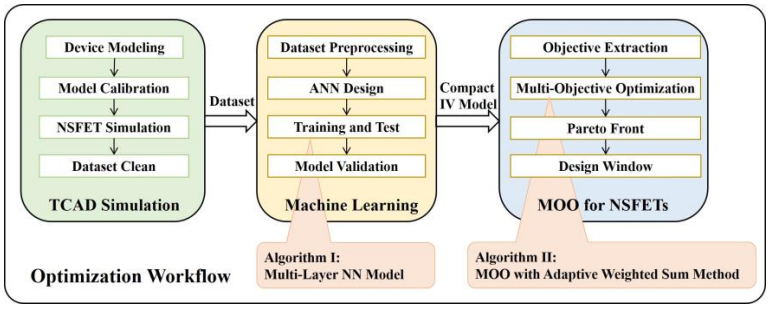


Fig. 1 Schematic of overall workflow for NSFET optimization.

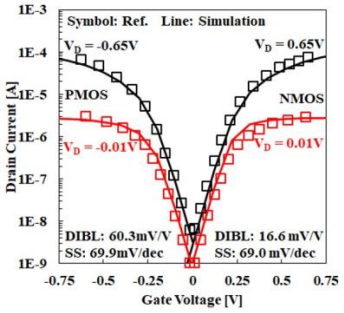


Fig.3 NSFETs calibration results of I_d-V_g curve with Ref. [2].

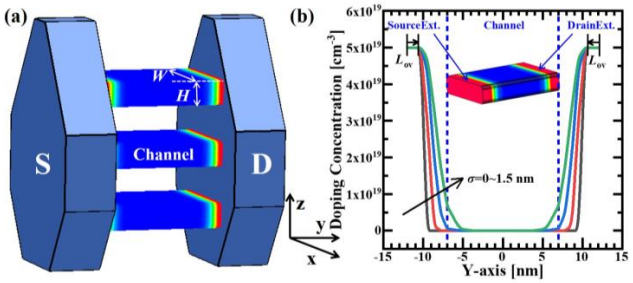


Fig.2 (a) Schematic of the three-layer-stacked NSFET model and (b) doping profile following various Gaussian Distribution in NSFETs.

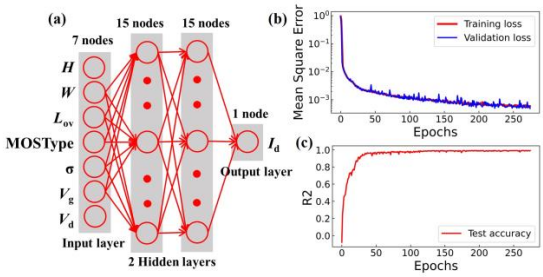


Fig. 4 (a) ANN trained in this work. (b) Loss function MSE on train set and validation set in training and (c) prediction accuracy R2 on the test set.

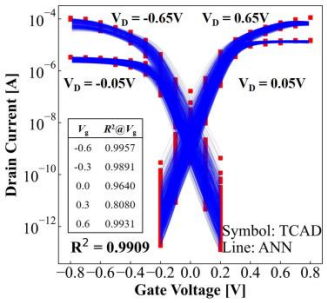


Fig. 5 Predicted and actual drain current versus gate voltage on the test set.

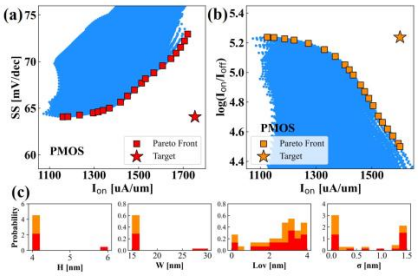


Fig. 6 Pareto Front of (a)SS versus Ion, (b)log(Ion/Ioff) versus Ion as two objectives for p-type NSFETs, and (c)their Pareto optimal solution distribution.

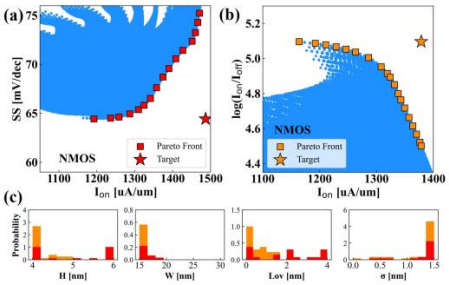


Fig. 7 Pareto Front of (a)SS versus Ion, (b)log(Ion/Ioff) versus Ion as two objectives for n-type NSFETs, and (c)their Pareto optimal solution distribution.

Table I Design Parameters		
Parameter	Calibration Value	Data Generation (Min, Max, Step)
H [nm]	5	(4, 6, 0.5)
W [nm]	20	(15, 30, 5)
Lov [nm]	1	(0, 4, 0.5)
σ [nm]	1.5	(0, 1.5, 0.5)
MOSType	P, N	P, N

Table II Optimal Design		
Parameter	PMOS	NMOS
H [nm]	4.0	4.0
W [nm]	15.0	15.0
Lov [nm]	3.0	0.0
σ [nm]	0.0	1.5

Table III Performance Comparison with IRDS Targets				
	PMOS	NMOS	2025Target	2028Target
L_g[nm]	14	14	14	14
V_{dd} [nm]	-0.65	0.65	0.65	0.65
SS[mV/dec]	64.8	61.6	72	75
DIBL[mV/V]	28.4	30	-	-
I_{on}/I_{off}	9.86E4	9.70E3	8.73E4	9.24E4
I_{on} [uA/um]	1429	1155	873	924
V_t [mV]	167	148	212	226

Reference

[1] N. Loubet et al., 2017 *VLSI*, pp. T230–T231.
[2] G. D. Yakimets et al., 2017 *IEDM*, pp. 20.4.1–20.4.4.
[3] Z. Stanojevic et al., 2015 *IEDM*, pp. 5–1.
[4] R. Hecht-Nielsen, *Neural networks for perception*, 1992, pp. 65–93.
[5] I. Y. Kim et al, *Structural and multidisciplinary optimization*, vol. 29, no. 2, pp. 149–158, 2005.

Ab Initio Quantum Transport Simulations of Monolayer GeS Nanoribbons

Mislav Matić and Mirko Poljak*

Computational Nanoelectronics Group
Faculty of Electrical Engineering and Computing, University of Zagreb, HR-10000 Zagreb, Croatia
*Corresponding author: mirko.poljak@fer.hr

Introduction. Monolayer GeS was recently examined along with hundreds of 2D materials as one of the most promising 2D materials for ultra-scaled FETs in [1], while a quantum transport study of sub-10 nm monolayer GeS FETs was reported in [2]. On the other hand, very little is known about GeS nanoribbons (GeSNRs) or GeSNR device performance, with the former being limited to electronic properties reported in [3]. Therefore, in this work we study electronic and transport properties of ultra-scaled GeSNRs and assess the ballistic performance of GeSNR FETs.

Methodology. Ultra-scaled armchair GeS nanoribbons passivated with H atoms (Fig. 1) are simulated using the DFT package Quantum Espresso [4] with Perdew-Burke-Ernzerhof generalized gradient approximation [5] (PBE-GGA) exchange-correlation (XC) functional. Maximally-localized Wannier functions (MLWFs) [6] are used to transform DFT Hamiltonians into a localized basis, which are then used to construct MLWF Hamiltonians of ~15 nm long nanoribbons of various widths (W). Our existing non-equilibrium Green's function (NEGF) quantum transport code is used for the calculation of geometry-dependent electronic and transport properties of GeSNRs. Contacts are assumed to be ideal, i.e. semi-infinite semiconducting material same as the channel, with Sancho-Rubio method [7] employed for the calculation of S/D contact self-energy matrices. Top-of-the-barrier (ToB) model [8] is used for ballistic simulations of the GeSNR MOSFET. Although simple, due to the inclusion of a full bandstructure, ToB is a dependable method for FETs with channel lengths > 15nm where direct S/D tunneling is negligible [8]. Gate oxide has EOT = 1 nm, and S/D doping is set at 0.001 molar fraction of the GeS areal density. For a meaningful comparison, in all devices we set V_{TH} as in the Intl. Roadmap for Devices and Systems (IRDS) at "3 nm" logic node [9], i.e. at 0.24 V, which results in an unrealistically low OFF-state current (I_{OFF}) of 0.87 nA/μm due to 60 mV/dec subthreshold slope and ideal gate control in the ToB model. The ballistic ON-state current (I_{ON}) is extracted at $V_{GS} = V_{DS} = 0.7$ V.

Results and discussion. Scaling down of nanoribbon width increases the bandgap (E_g) from 1.73 eV for $W = 3.70$ nm to 2.24 eV for $W = 0.76$ nm as shown in Fig. 2. The bandstructure plots for GeSNRs of various widths, reported in Fig. 3, show that E_g of wider GeSNRs is indirect while scaling down the nanoribbon width below $W = 2.23$ nm transitions GeSNRs into a direct semiconductor which agrees with the results in [3]. As shown in Fig. 4, ON-state current decreases monotonically while scaling down the GeSNR width, from $I_{ON} = 1.20$ mA/μm for $W = 3.70$ nm to $I_{ON} = 0.65$ mA/μm for $W = 0.76$ nm. The performance of GeSNR FETs is directly related to the channel bandstructure along nanoribbon transport direction, as reported in Fig. 3, since mobile charge density depends on density of states (DOS) and transmission function determines transport probability for each conducting mode. Subbands nearest to the conduction band minimum (CBM) have the highest influence on the ON-state current in *n*-channel FETs. For nanoribbons with $W = 2.97$ nm and $W = 3.70$ nm, I_{ON} values are approximately the same due to bandstructure similarity near the CBM for both GeSNRs. To further clarify the I_{ON} - W curve, transmission and DOS are plotted in the ~200 meV energy range from the CBM, with CBM shifted to 0 eV for comparison as shown in Fig. 5. Nanoribbons with $W = 3.70$ nm and $W = 2.97$ nm show DOS and transmission characteristics that match almost perfectly for energies up to 0.11 eV away from the CBM, where the contribution to the current is the highest, thus resulting in the same I_{ON}. The 2.23 nm-wide GeSNR presents a transitional nanoribbon because lowest sub-bands, nearest to the CBM, start moving away from the CBM as W decreases further (compare e.g. Fig. 3c and d). This subband shift directly translates to the significant I_{ON} drop while scaling the GeSNR width from 2.97 nm to 2.23 nm observed in Fig. 4. For GeSNRs with $W < 2.23$ nm the subband with lower curvature, i.e. higher effective mass, becomes the lowest and dominant subband. This property, alongside the lower number of available bands/modes due to the lower number of available orbitals in narrower GeSNRs, further decreases I_{ON} in ultra-scaled GeSNR FETs. None of the analyzed devices fulfills IRDS specification for I_{ON}, but the performance can be improved by EOT, doping, etc. optimization, which is beyond the scope of this work. On the other hand, I_{ON}/I_{OFF} reaches ~7×10^5 even in the worst case, indicating good switching capabilities of GeSNR FETs for logic applications.

Conclusion. We used NEGF and MLWF Hamiltonians to study the electronic, transport and ballistic device properties of sub-4 nm-wide and ~15 nm-long GeSNRs. While ultra-scaled GeSNR FETs exhibit good I_{ON}/I_{OFF} of at least ~7×10^5, indicating good logic switching performance, they also offer modest ballistic I_{ON} values of up to ~1.20 mA/μm, for GeSNRs with $W < 4$ nm widths. Further performance improvement is possible, but future work must also consider carrier scattering and dissipative transport for a realistic assessment of GeS nanoribbon devices.

References **[1]** C. Klinkert, Á. Szabó, C. Stieger, D. Campi, N. Marzari, and M. Luisier, "2-D Materials for Ultrascaled Field-Effect Transistors: One Hundred Candidates under the *Ab Initio* Microscope," *ACS Nano*, vol. 14, no. 7, pp. 8605–8615, Jul. 2020. **[2]** Y. Ding *et al.*, "High-Performance Ballistic Quantum Transport of Sub-10 nm Monolayer GeS Field-Effect Transistors," *ACS Appl. Electron. Mater.*, vol. 3, no. 3, pp. 1151–1161, Mar. 2021. **[3]** R. Li, H. Cao, and J. Dong, "Electronic properties of group-IV monochalcogenide nanoribbons: Studied from first-principles calculations," *Phys. Lett. A*, vol. 381, no. 44, pp. 3747–3753, Nov. 2017. **[4]** P. Giannozzi *et al.*, "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials," *J. Phys. Condens. Matter*, vol. 21, no. 39, Art. no. 39, Sep. 2009. **[5]** J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Phys. Rev. Lett.*, vol. 77, no. 18, Art. no. 18, Oct. 1996. **[6]** N. Marzari and D. Vanderbilt, "Maximally localized generalized Wannier functions for composite energy bands," *Phys. Rev. B*, vol. 56, no. 20, Art. no. 20, Nov. 1997. **[7]** M. P. L. Sancho, J. M. L. Sancho, and J. Rubio, "Quick iterative scheme for the calculation of transfer matrices: application to Mo (100)," *J. Phys. F Met. Phys.*, vol. 14, no. 5, Art. no. 5, May 1984. **[8]** A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom, "Theory of Ballistic Nanotransistors," *IEEE Trans. Electron Devices*, vol. 50, no. 9, p. 13, 2003. **[9]** IEEE Intl. Roadmap for Devices and Systems (IRDS), https://irds.ieee.org

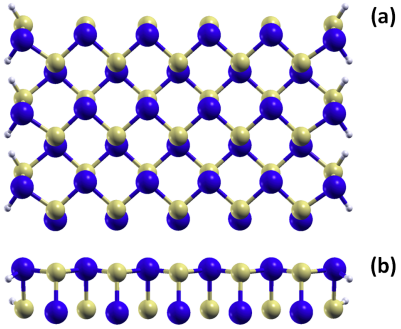


Figure 1. (a) Top and (b) side view of monolayer GeS armchair nanoribbon. Edges are H-terminated.

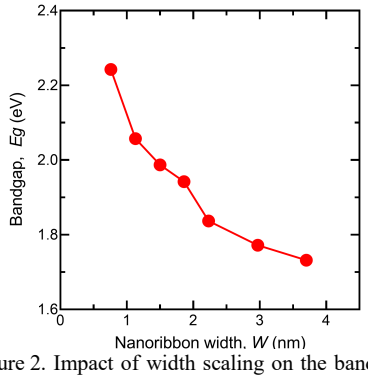


Figure 2. Impact of width scaling on the bandgap of GeSNRs.

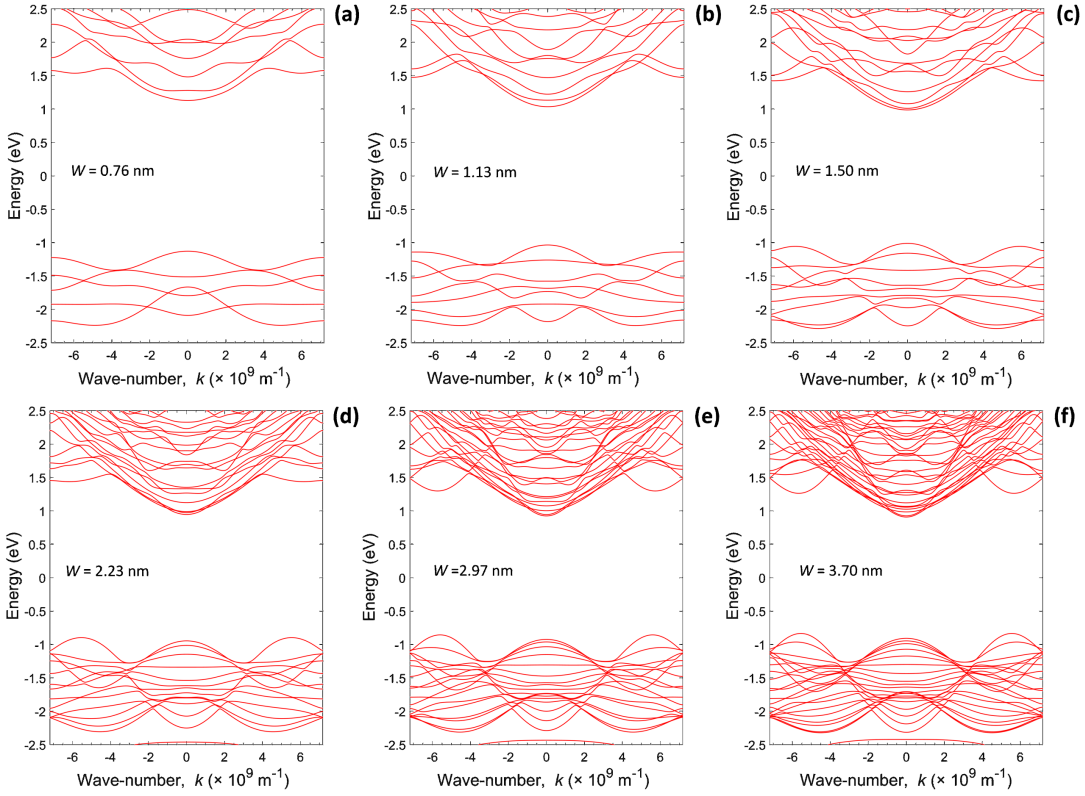


Figure 3. Bandstructure comparison of GeSNRs with the widths of (a) $W = 0.76$ nm, (b) $W = 1.13$ nm, (c) $W = 1.50$ nm, (d) $W = 2.23$ nm, (e) $W = 2.97$ nm, and (f) $W = 3.70$ nm.

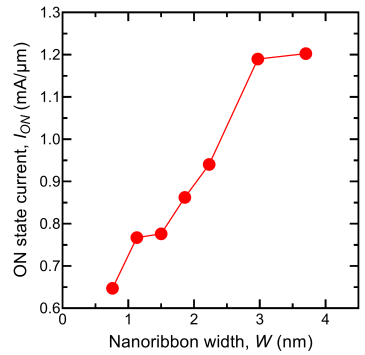


Figure 4. ON-state current vs. width down-scaling in 15 nm-channel GeSNR MOSFETs.

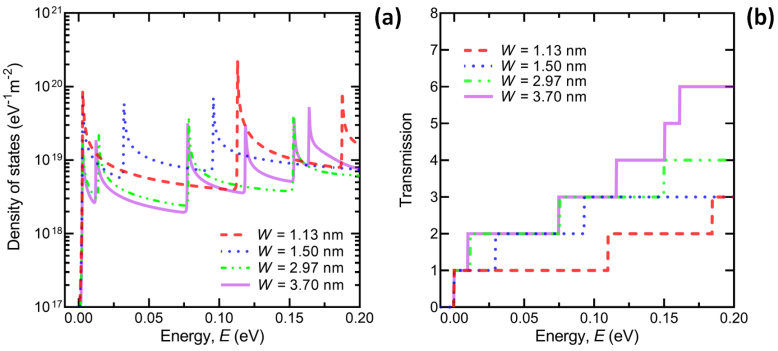


Figure 5. Comparison of (a) density of states and (b) transmission for GeSNRs with the following widths: $W = 1.50$ nm (dotted blue line), $W = 1.86$ nm (dashed red line), $W = 2.97$ nm (dot-dot-dashed green line), and $W = 3.70$ nm (purple line).

Acceleration of Semiconductor Device Simulation Using Compact Charge Model

Kwang-Woon Lee and Sung-Min Hong
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology,
123 Cheomdan-gwagiro (Oryong-dong), Buk-gu, Gwangju, 61005, Republic of Korea
E-mail: smhong@gist.ac.kr

Abstract

In this work, we propose a method to get an initial guess for the semiconductor device simulation with a compact charge model. By using the obtained initial guess, we can perform the device simulation directly at the target bias condition without a time-consuming bias ramping process. In order to verify our method, rectangular Gate-All-Around (GAA) MOSFETs having a long channel length are considered. Results clearly show that the device simulation can be accelerated through our method.

Introduction

In the semiconductor device simulation, the target bias condition has to be reached through the bias ramping process, which is typically time consuming. Therefore, when a good initial guess for the target bias condition is available, the computational time can be significantly reduced. Our group has reported a method to generate the initial guess through a trained deep neural network [1,2,3,4]. The efficiency of the proposed method has been demonstrated with several numerical examples.

In this study, instead of a deep neural network which must be trained before the inference, we use a compact charge model to predict an initial guess. The compact charge model for a 2D MOS structure is numerically solved with the 1D continuity equation. The solution is used to generate an initial guess. Rectangular GAA MOSFETs shown in Fig. 1 are simulated by using the proposed method.

Results for 2D Rectangular MOS structure

In this work, we use a compact charge model for 2D MOS structures [5]. It is expressed as:

$$V_G - \Phi_{MS} + \frac{Q_d}{P\langle C_{ins}\rangle_s} + \frac{Q_e}{P\langle C_{ins}\rangle_s} \approx \langle\phi\rangle_s\,, \qquad (1)$$

$$\langle\phi\rangle_s \approx V + V_T \log\left(\frac{Q_e^2 + 2(1-\alpha_e)Q_eQ_d}{2q\epsilon P^2 V_T n_{int}\left(1-\beta\exp\left(\frac{A^*}{P^2}\frac{2\alpha_eQ_e+Q_d}{2\epsilon V_T}\right)\right)}\right), \qquad (2)$$

where Q_e is the integrated electorn charge, V is the electron quasi-Fermi potetnial, α_e is a noramlized distance between the electron centroid and the interface, and β is a correction factor to consider an effect of a cross-section. All other symbols have conventional meanings and more details are described in [5]. Furthermore, a generalized coordinate, ψ, is introduced, and it will be used for predicting the initial guess.

A procedure to predict the initial guess for 2D rectangular GAA MOS structures is as follows. First, parameters for calculating the compact model are extracted, and (1) and (2) are solved at the target gate voltage. Then, Q_e, $\langle\phi\rangle_s$, and $\langle\phi\rangle_X$ can be obtained at the targe gate voltage. Here, $\langle\phi\rangle_X$ is an average of ϕ over the cross-section with $\nabla^2\psi$ as a weighting factor [5]. Next, by using $\langle\phi\rangle_s$ and $\langle\phi\rangle_X$, we predict the initial potential, $\phi_{initial}$, for the semiconductor region at the target bias condition as follows:

$$\phi_{initial} = max(\langle\phi_n'\rangle_s(\psi - \Psi) + \langle\phi\rangle_s, \langle\phi\rangle_X), \qquad (3)$$

where Ψ is the value of ψ at the interface and $\langle\phi_n'\rangle_s = -\frac{Q_e+Q_d}{\epsilon P}$. Through this procedure, $\phi_{initial}$ can be obtained. Figure 2 shows differences between $\phi_{initial}$ and the final converged potential. Three different aspect ratios are considered. The maximum difference between the initial potential and the converged one is about 77 mV. By using $\phi_{initial}$, at most 7 Newton iterations are needed at the target bias condition without any bias ramping for solving the Poisson equation of MOS structures in Fig 2.

Results for 3D Rectangular GAA MOSFET

The procedure for 2D MOS structures is extended to 3D MOSFETs. In order to obtain the initial guess for the drift-diffusion model, we solve the 1D electron continuity equation with (1) and (2). The 1D electron continuity equation is expressed as follows:

$$\frac{d}{dz}\left(\mu Q_e\frac{dV}{dz}\right) = 0\,, \qquad (4)$$

where μ is the electron mobility. In this preliminary work, a constant mobility is assumed. By solving (1), (2), and (4) together at the target bias point, we can get Q_e and V along the channel direction. By using the calculated Q_e and V, the initial potential profiles can be predicted. Also, from the initial potential, the inital guess for the electron density can be obtained. These initial guesses are used to calculate the drift-diffusion model.

In Fig. 3, results for rectangular GAA MOSFETs are shown. In this study, we use $1\,\mu m$-long channel with $10^{16}cm^{-3}$ channel doping concentraion and $10^{20}cm^{-3}$ source/drain doping. The aspect ratio of the cross-section used in Fig. 3 is 1:3 (W = 18nm and H = 6nm). Figure 3 shows the initial Q_e and V along the z-direction compared with the final converged results at several bias conditions. Differences of V between initial guesses and final results are also shown and the number of Newton iterations needed for solving drift-diffusion model at the target bias condtion without any ramping process is found for each case. In these examples, at most 6 Newton iterations are needed to obtain the self-consistent solution. It is much more efficient than the conventional bias ramping method.

Conclusions

In conclusion, by using the compact charge model, the initial guess for the semiconductor device simulation is obtained appropriately. With the obtained initial guess, the device simulation is performed directly at the target bias point without any bias ramping. By adopting the proposed method, the number of Newton iterations to get the converged solution can be significantly reduced.

References

[1] S.-C. Han and S.-M. Hong, SISPAD, 2019.

[2] S.-C. Han, J. Choi, and S.-M. Hong, SISPAD, 2020.
[3] S.-C. Han, J. Choi, and S.-M. Hong, SISPAD, 2021.
[4] S.-C. Han, J. Choi, and S.-M. Hong, TED, vol. 68, 2021.
[5] K.-W. Lee and S.-M. Hong, TED, accepted.

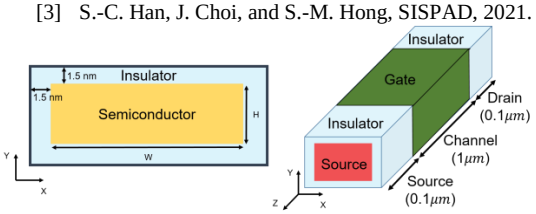


Figure 1. Cross-section of a rectangular GAA MOSFET and its 3D structure. The aspect ratio used in this work means a ratio of W and H. Thickness of the insulator is 1.5 nm. Channel length is 1 μm and the source/drain region is 0.1 μm-long. The p-type doping concentration of the channel region is $10^{16} cm^{-3}$ and the source/drain n-type doping is $10^{20} cm^{-3}$ in this work.

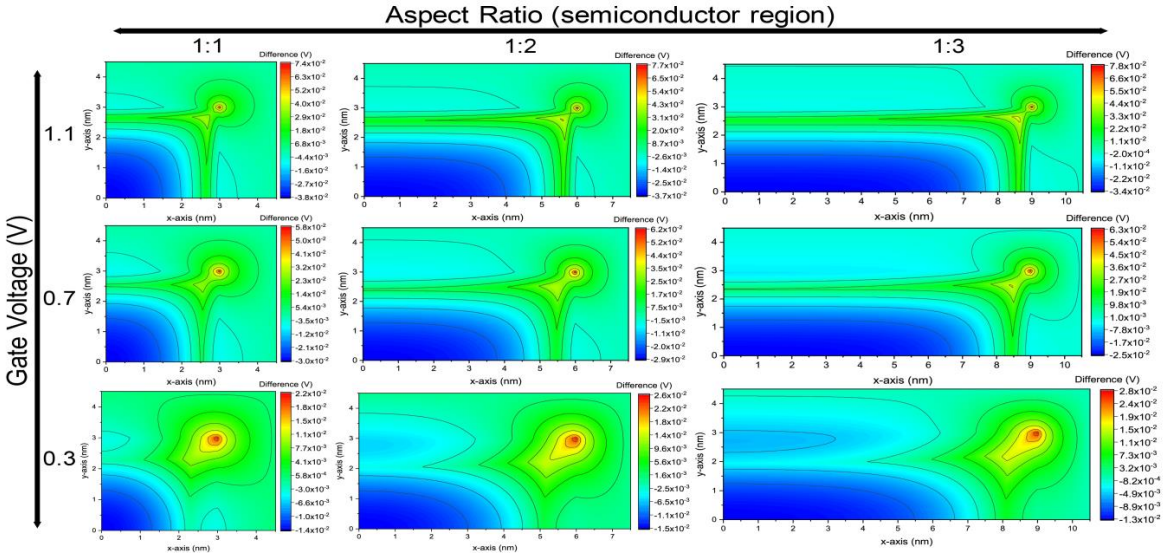


Figure 2. Difference between the initial guess and the converged solution. Three aspect ratios and three gate voltages are considered. Only a quarter of the cross-section is shown. In these examples, the maximum difference is about 77 mV.

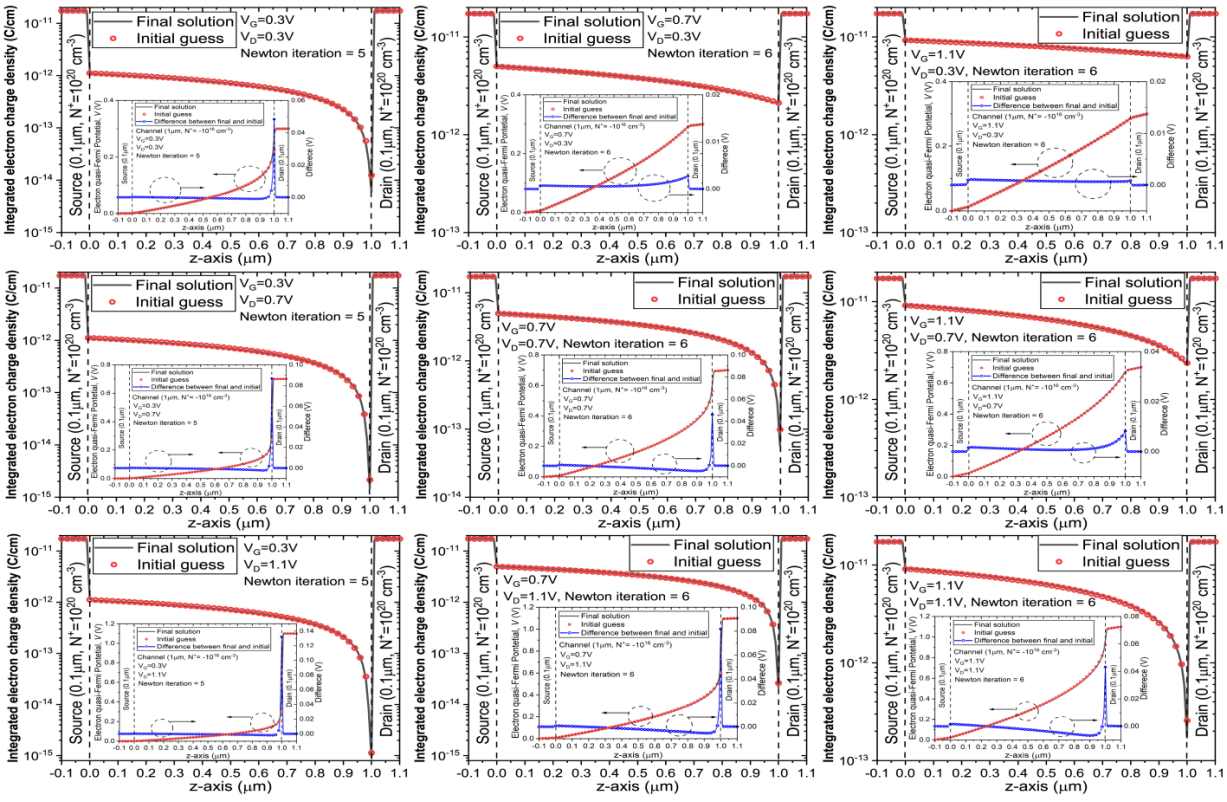


Figure 3. Initial guess of Q_e and V along the z-direction compared with final results at several bias conditions. The difference of V between initial guess and final result is also shown. Through the method proposed in this work, at most 6 Newton iterations are needed for solving the drift-diffusion model.

Atomic-scale study of silane and hydrogen adsorptions competition during Si epitaxy

Laureline Treps*, Jing Li*, Benoît Sklénard*†
*Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France
†E-mail: benoit.sklenard@cea.fr

***Abstract*—Epitaxy by Chemical Vapor Deposition (CVD) is one of the mainly used method of surface growth of Si alloys. An important industrial challenge of today is to adapt this process to low-temperature conditions. In this work, we study the generic adsorption of silane as a precursor for epitaxy on Si(001) surface by means of Density Functional Theory (DFT). Emphasis is placed on the adsorption with different hydrogen coverages and the limiting effect of hydrogen desorption. A partial coverage of the surface is shown to be thermodynamically and kinetically favorable. The crutial point to improve the process would be to maintain the surface with this partial coverage in hydrogen. Another limiting element would be to bring enough energy for hydrogen desorption while avoiding silane desorption which needs only 0.2 eV more.**

I. Introduction

The continuous scaling of Complementary Metal Oxide Semiconductor (CMOS) devices has brought the development of disruptive technologies such as 3D integrations where stacked layers are processed sequentially. Such integrations impose thermal budget constraints to the top tier fabrication process (around 500°C) as to avoid any degradation of bottom tier devices [1]. Therefore, process steps such as chemical vapor deposition (CVD) epitaxy have to be adapted to fulfill these temperature contraints [2], [3]. At low temperatures, the growth rate with standard precursors such as SiH_4 or SiH_2Cl_2 is regulated by the H surface coverage. However the detailed adsorption and desorption phenomena remain poorly understood. In this work, we study the impact of H coverage on the adsorption of SiH_4 on a Si(001) surface by means of *ab initio* simulations and provide a detailed reaction mechanism of the growth process in the low temperature regime.

II. Computational methods

All calculations based on Density Functional Theory (DFT) have been carried out using the Vienna Ab initio Simulation Package (VASP) code [4], [5] with PAW atomic datasets including $3s^2 3p^2$ valence electrons for Si. The structures are relaxed using Perdew-Burke-Ernzerhof (PBE) exchange-correlation functionals [6] with periodic boundary conditions and with D3 dispersion correction of Grimme *et al.* to describe Van der Waals (VdW) interactions [7]. We employed an energy cutoff of 250 eV for the plane-wave basis and a $2 \times 2 \times 1$ k-mesh.

During a chemical reaction, reactant and products are separated by an energy barrier which can be investigated. This energy barrier is defined by the energy difference between the energy of the reactant and the maximum of the reaction path which is called a transition state (TS). A TS is a saddle point of the potential energy landscape and can be found with the "nudged elastic band" (NEB) method [8]. It is possible to climb uphill and converge to the TS with the adapted CI-NEB method [9].

In this study, we consider a 4×2 Si(001) slabs with 8 layers. The bottom surface is passivated with dihydride termination. The top surface corresponds to the c(4×2) reconstructed surface with a coverage θ=0 ML, this model contains 32 hydrogen atoms and 128 silicon atoms.

III. Results and discussions

A. Si(001) surface in epitaxy temperature and hydrogen pressure conditions

The coverage of the surface, the adsorption and desorption of hydrogen are studied in various studies on smaller surface. These calculations show that the coverage for low temperature epitaxy (≤850°C) remains between θ=0.5 ML (128 Si atoms and 40 H atoms) and θ=1 ML (128 Si atoms and 48 H atoms) depending on temperature and H_2 partial pressure and that the surface dimers are maintained for coverage values of 1 ML and lower. The experimental value of hydrogen desorption is evaluated at 2.5 eV. The theoretical value calculated by Tsatsoulis *et al.* [10] with PBE is 1.95 eV with a 4×2 Si(001) slab while our value is 2.19 eV. This step is crucial to understand what the surface looks like for the surface construction with epitaxy.

B. Silane adsorption on Si(001) surface

The construction of Si(001) crystal surface with the epitaxy process can be initiated with various precursors such as SiH_2Cl_2, Si_2H_6, or SiH_4. For a generic example we study the silane adsorption over this surface as presented in the following reaction.

$$SiH_4(g) \longrightarrow SiH_3^* + H^*$$

Different parameters can be taken into account for this reaction and have been studied separately: the hydrogen coverage [11], the adsorption mode of molecules (intradimer, interdimer, inter-row) [12], and the supercell size [13]. The first one has been studied for the adsorption of SiH_3 only while the two others have been studied for the same reaction. The calculated adsorption energy is evaluated to 0.26 eV for intra adsorption, 0.31 eV for inter adsorption, and 0.81 eV for inter-row. The intra adsorption on a larger cell has a computed value of 0.42 eV. The experimental values are under 0.20 eV [14] for the intra adsorption and under 0.44 eV [15] for the inter adsorption.

In our study, we combined the three parameters to have a global understanding of the phenomenon. As an example, we present here the inter-dimer adsorption of silane on a large cell model and with different hydrogen coverages (0 and 0.875 ML). The results are presented in Fig. 1, the initial and the final configurations can be seen for each coverage as well as the reaction paths. The reaction energy E_r (difference between the energy of the final state TS and the initial state IS) of the three different paths already highlights an important difference due

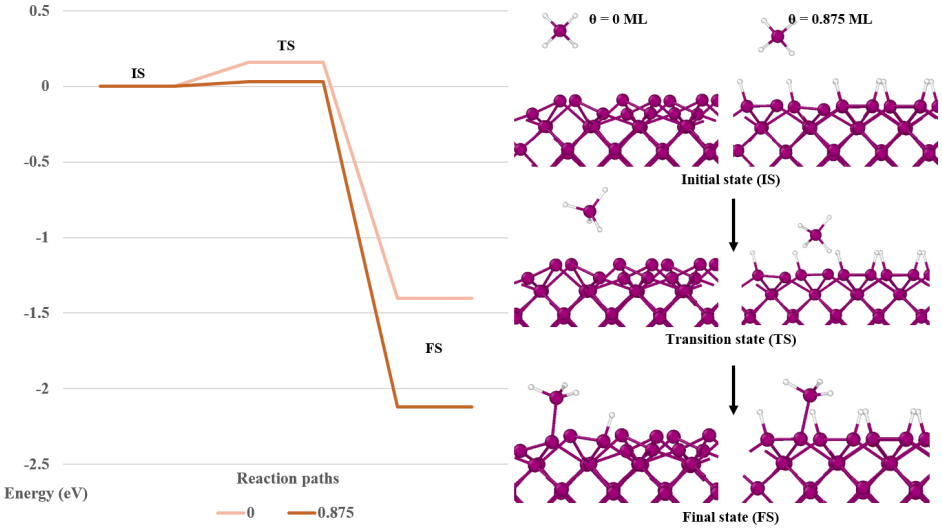


Fig. 1. Reaction path energies for silane adsorption on two dimers of Si(001) surface for different hydrogen coverages (θ = 0 ML and 0.875 ML). The energies (in eV) of the initial states are set at 0 eV for reading commodities. Hydrogen atoms are represented in white and silicon atoms in purple.

to the coverage: the high coverages (0.875 ML) have reaction energies of -2.2 eV respectively while the empty surface has a reaction energy of -1.4 eV. A partial coverage of the surface (not all the dimers with hydrogen) seems favorable to the adsorption of silane. The activation energy E_a (difference between the transition state energy TS and the initial state IS) of the two different paths show a small difference. The partial coverages of the surface (θ = 0.875 ML), has an activation energy of 0.03 eV respectively while the empty surface has an activation energy of 0.16 eV. The energetic barrier is lower with partial coverage of the surface. These energies show that the adsorption of silane is thermodynamically and kinetically more favorable with a partial coverage of the surface. The desorption of hydrogen is crutial to continue the construction of the silicon surface. This reaction is in competition with silane desorption. The calculated desorption energy of hydrogen is 2.4 eV when neighbouring dimer have hydrogen adsorbed on it. While if neigbouring dimer are empty or have some hydrogen adsorbed, the desorption energy of hydrogen is between 2.1 and 2.2 eV.The desorption energy of silane is 2.2 eV for 0.875 ML coverage and 1.6 eV with a surface free of hydrogen. The difference between hydrogen and silane desorption is smaller (0.2 eV) for higher surface coverage in hydrogen than for low hydrogen coverage (0.5 eV).

IV. Conclusions

To conclude, we highlighted that the initial coverage of the surface determines the energetic cost of the epitaxy process. A partial hydrogen coverage of the Si(001) surface will favor thermodynamically and kinetically the adsorption of silane. These partial coverage of Si(001) clearly correspond to the low-temperature epitaxy conditions, and the adsorption of silane on this surface will clearly be favored in these conditions. Also for low-temperature conditions, the difference in energy desorption between silane and hydrogen is 0.2 eV in favor of hydrogen desorption. The epitaxy, the growth of silicon surface, must be favored.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871813 MUNDFAB and was performed using HPC resources from GENCI-IDRIS (Grant 2021-A0110911995).

References

[1] L. Brunet, S. Reboh, T. Januel, X. Garros, T. M. Frutuoso, M. Cassé, B. Sklénard, L. Brévard, M. Ribotta, A. Magalhaes-Lucas, J. Kanyandekwe, J.-M. Hartmann, F. Milesi, F. Mazen, P. Acosta-Alba, S. Kerdilès, A. Tavernier, V. Loup, C. Morales, V. Larrey, F. Fournel, L. L. Van-Jodin, S. Leforestier, E. Rolland, G. Romano, G. Gaudin, J. Lugo, J. Lacord, S. Maitrejean, J. Arcamone, P. Batude, I. Radu, C. Fenouillet-Beranger, and F. Andrieu in *2021 Symposium on VLSI Technology*, pp. 1–2, 2021.
[2] C. Fenouillet-Beranger, L. Brunet, P. Batude, L. Brevard, X. Garros, M. Cassé, J. Lacord, B. Sklenard, P. Acosta-Alba, S. Kerdilès, A. Tavernier, C. Vizioz, P. Besson, R. Gassilloud, J.-M. Pedini, J. Kanyandekwe, F. Mazen, A. Magalhaes-Lucas, C. Cavalcante, D. Bosch, M. Ribotta, V. Lapras, M. Vinet, F. Andrieu, and J. Arcamone *IEEE Trans Electron Devices*, vol. 68, no. 7, pp. 3142–3148, 2021.
[3] C. Porret, A. Y. Hikavyy, J. F. G. Granados, S. Baudot, A. Vohra, B. Kunert, B. Douhard, J. Bogdanowicz, M. Schaekers, D. Kohen, J. Margetis, J. Tolle, L. Lima, A. Sammak, G. Scappucci, E. Rosseel, R. Langer, and R. Loo *ECS Transactions*, vol. 86, pp. 163–175, jul 2018.
[4] G. Kresse and J. Furthmüller *Phys. Rev. B*, vol. 54, p. 11169–11186, Oct 1996.
[5] G. Kresse and D. Joubert *Phys. Rev. B*, vol. 59, p. 1758–1775, Jan 1999.
[6] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple (vol 77, pg 3865, 1996)," 1997.
[7] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg *J. Chem. Phys.*, vol. 132, no. 15, p. 154104, 2010.
[8] G. Henkelman, B. P. Uberuaga, and H. Jónsson *J. Chem. Phys.*, vol. 113, no. 22, pp. 9901–9904, 2000.
[9] G. Henkelman, B. P. Uberuaga, and H. Jónsson *J. Chem. Phys.*, vol. 113, no. 22, pp. 9901–9904, 2000.
[10] T. Tsatsoulis, S. Sakong, A. Groß, and A. Grüneis *J. Chem. Phys.*, vol. 149, no. 24, p. 244105, 2018.
[11] T. N. Le, P. Raghunath, L. K. Huynh, and M.-C. Lin *Appl. Surf. Sci.*, vol. 387, pp. 546–556, 2016.
[12] J. Shi, E. S. Tok, and H. C. Kang *J. Chem. Phys.*, vol. 127, no. 16, p. 164713, 2007.
[13] *Appl. Surf. Sci.*, vol. 496, p. 143728, 2019.
[14] D. F. Kavulak, H. L. Abbott, and I. Harrison *J. Phys. Chem. B*, vol. 109, no. 2, pp. 685–688, 2005.
[15] M. Fehrenbacher, J. Spitzmüller, M. Pitter, H. R. H. Rauscher, and R. J. B. R. J. Behm *Jpn. J. Appl. Phys.*, vol. 36, no. 6S, p. 3804, 1997.

Automatic Grid Refinement for Thin Material Layer Etching in Process TCAD Simulations

Christoph Lenz*, Paul Manstetten†, Andreas Hössinger‡, and Josef Weinbub*
*Christian Doppler Laboratory for High Performance TCAD at the
†Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, 1040 Wien, Austria
‡Silvaco Europe Ltd., Compass Point, St Ives, Cambridge, PE27 5JL, United Kingdom
Email: lenz@iue.tuwien.ac.at

***Abstract*—The utilization of thin material layers is common in modern semiconductor device fabrication. Subsequent etching steps require an accurate modeling of these thin layers. Although level-set based process TCAD simulations are capable of representing flat thin material layers with sub-grid accuracy, topographical changes during etching processes expose the low underlying grid resolution, which leads to detrimental artifacts. We present a novel algorithm that analyzes the thickness of all material layers and suggests a refined target resolution for local regions of thin layers affected by the etching process. This allows to accurately represent topographical changes in thin layers without refining unaffected regions of the domain. We simulate the fabrication of a light-emitting diode device, where our algorithm is used to automatically predict the optimal resolution for all etched material layers. Our algorithm selects efficient refinement factors to obtain the target resolutions with locally employed hierarchical grids.**

Introduction The fabrication processes of many modern semiconductor devices – like light-emitting diodes (LEDs) or three-dimensional NAND *staircase* flash memories – include process steps in which parts of thin material layers are etched away [1], [2]. Such fabrication processes can be simulated by process technology computer-aided design (TCAD) simulations employing the level-set method. The topography of the wafer is described by a function ϕ in the domain Ω, where the zero level-set of ϕ is defined as:

$$\{x \in \Omega \mid \phi(x) = 0\}, \tag{1}$$

which is identical to the wafer surface. The level-set function is discretized on a regular grid with grid resolution Δx. Furthermore, all material layers within the simulated structure are represented using individual level-set functions [3].

Several etching processes in a level-set based simulation framework use Boolean operations of level-set functions to obtain the final status of the structure after the etching process [3]: The to-be-removed parts of material layers from the wafer (by etching) is also represented by a level-set function. The relative complement (Boolean operation) of this level-set function and all material layers affected by the etching process is calculated within the etching algorithm and represents the state of the wafer after the etching process is completed (see Figure 1).

The level-set method is capable of representing flat thin material layers (nm regime) with a very coarse (µm) resolution. However, when parts of these thin material layers are removed by a Boolean operation the low lateral resolution of the underlying simulation domain gets exposed and the material layers meld into each other, introducing artifacts in the resulting representation of the structure. To illustrate this effect, Figure 2 shows three different grid resolutions (Δx) from 1/4 to 6 grid points in vertical direction (thickness of each material layer). Increasing the resolution in the whole simulation domain is prohibitive due to severe hits on overall simulation performance. Hierarchical grids can be used to enable a locally increased resolution. A simulation employing hierarchical grids consist of a base-grid that spans the entire simulation domain and several nested sub-grids that cover areas of interest (features) with a higher resolution [4]. A feature detection algorithm based on the geometrical features of the topography, like the surface curvature, is commonly used to guide the refinements during the evolution of the topography [4]. Such an approach is well suited for detecting emerging features in the topography. However, it detects features only after a Boolean operation has been performed. To maintain the accuracy of the representation after a Boolean operation has been applied, the resolution refinement has to be performed before the Boolean operation is applied.

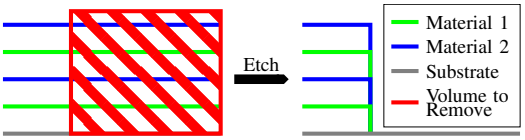
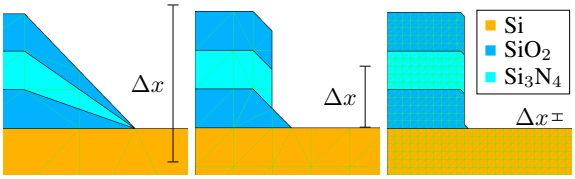


Fig. 1: Illustration of four stacked material layers on a wafer and a simulated etching process using a Boolean operation.



(a) $\Delta x = 0.75$µm (b) $\Delta x = 0.25$µm (c) $\Delta x = 0.005$µm

Fig. 2: Thin material layers with a thickness of 0.25µm after a Boolean operation with different base-grid resolutions.

In this work, we present an algorithm for Boolean operations on hierarchical grids that automatically analyzes the thickness of material layers which are affected by a Boolean operation and calculates a minimum refinement level (before the operation is applied) to obtain a sufficiently accurate geometric representation.

The proposed algorithm has been implemented into Silvaco's Victory Process simulator [5].

Method To determine the minimal required grid resolution (Δx_{tar}) to accurately represent a thin material layer affected by a Boolean operation. We have to take into account a predefined minimal number of grid points to represent a material layer (N_{min}) as well as the distance to the closest other material layer affected by the Boolean operation (d_{closest}), it follows for a given grid point:

$$\frac{d_{\text{closest}}}{N_{\text{min}}} = \Delta x_{\text{tar}}. \tag{2}$$

This information combined with the grid resolution of the current sub-grid (Δx_{cur}) and a user-supplied refinement factor (F_{ref}) is subsequently used to calculate the required number of refinement levels (L_{ref}):

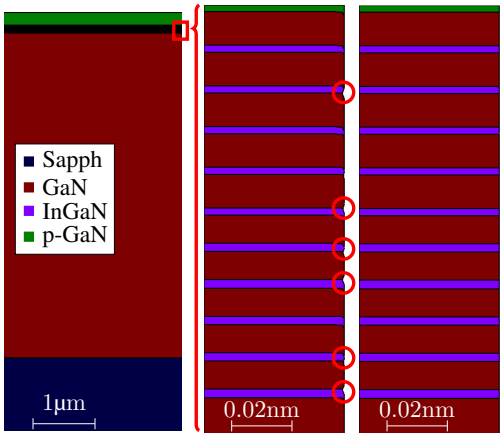
$$\left\lceil \frac{\log(\frac{\Delta x_{\text{curr}}}{\Delta x_{\text{tar}}})}{\log(F_{\text{ref}})} \right\rceil = L_{\text{ref}}. \tag{3}$$

The availability of the target resolution (Δx_{tar}) – which is not available using a topography-based feature detection – allows our algorithm to deviate from a fixed refinement factor for the newly created sub-grids.

A Boolean operation then starts with the calculation of the level-set function χ that describes the material that is removed on the base-grid. After χ has been calculated, all material layers affected by the Boolean operation are determined. Next, the grid points at the intersection of the affected material layers and the level-set function χ are determined. At these intersection points, the distance form the material layer that causes the intersection to all other material layers is calculated. The distance to the closest material layer in normal direction is checked if it fulfills $\frac{d_{\text{closest}}}{N_{\text{min}}} \leq \Delta x_{\text{tar}}$: If not, the grid point is marked for refinement. After all intersection points have been checked, a hierarchical grid placement algorithm is initiated. The level-set function χ is recalculated where a higher resolution is now available. The above procedure is repeated on the new sub-grids until a predefined number of grid levels is reached. Afterwards (3) is used to determine the refinement level of the final sub-grids to accurately resolve the thinnest material layers.

Results For evaluation, we simulate the fabrication of an LED pixel of an LED array [2], [6]: For that simulation we use a base-grid resolution of $0.125\mu\text{m}$, $N_{\text{min}} = 6$, $F_{\text{ref}} = 4$, and require a minimum of two grid refinement levels (4-4), to guarantee an accurate representation of corners in the entire domain.

On a (0001) sapphire substrate a $1.9\mu\text{m}$ thick GaN layer is grown, afterwards 10 alternating layers with different thickness of InGaN and GaN with a total height of 117.5nm are grown on top of the GaN. The thinnest material layer is a 3nm thick InGaN layer. Afterwards, a 210nm thick p-GaN cap layer is grown on top of the structure [2], [6]. To create a singular LED pixel with a diameter of $75\mu\text{m}$ the excess material is etched away using the here presented algorithm. It follows from equation 3 that the finest sub-grid has to have a 256 times finer resolution to accurately resolve the thinnest material layers. Our algorithm constructs this refinement by adding one additional sub-grid with a 16 times finer resolution (4-4-16). We also assess the option to apply the thin layer specific refinement in smaller steps rather than in one big step. Figure 3 shows the LED device after etching through the thin material layers (active region). Figure 3b shows that choosing a smaller N_{min} (resulting in 3 grid levels 4-4-4) is not enough to accurately represent the thin layers after the etching process (see visible kinks between the material layers). Figure 3c shows the results using a 4-4-16 and a 4-4-4-4 refinement which produce the same final topography. Table I shows the run time of our algorithm when it is allowed to deviate from a fixed refinement factor: Our algorithm has a two times faster simulation run time when it utilizes its ability to deviate from a fixed refinement factor. Furthermore, our algorithm has a three times faster run time than an algorithm based on geometrical features only.



(a) LED device (b) 4-4-4 (c) 4-4-16 / 4-4-4-4

Fig. 3: LED device: (a) entire device, (b) active region with 3 grid levels (fixed refinement factor) – red circles highlight kinks resulting from the too low resolution –, (c) active region with 4 grid levels.

Feature detection method	Refinement factors	Run time
Our method	4-4-16	4 min 22 s
Our method	4-4-4-4	8 min 45 s
Geometrical	4-4-4-4	11 min 45 s

TABLE I: Etching simulation run times for different refinement level configurations. (Intel Xeon E5-2680v2)

Acknowledgment

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged.

References

[1] P. Hong *et al.*, *IEEE Access*, vol. 8, pp. 140 054–140 061, 2020.
[2] X. Zhou *et al.*, *Prog. Quant. Electron.*, vol. 71, p. 100263, 2020.
[3] O. Ertl, "Numerical Methods for Topography Simulation," Ph.D. dissertation, TU Wien, 2010.
[4] C. Lenz *et al.*, *Solid State Electron.*, vol. 191, p. 108258, 2022.
[5] Silvaco, "Victory Process," 2022. [Online]. Available: www.silvaco.com/tcad/victory-process-3d/
[6] S. Zhang *et al.*, *IEEE Photonics J.*, vol. 4, pp. 1639–1646, 2012.

Comparative Analysis of NBTI Modeling Frameworks – BAT and Comphy

Aseer Israr Ansari, Nilotpal Choudhury, Narendra Parihar and Souvik Mahapatra*

Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, 400076, India
* Phone: +91-222-572-0408, E-mail: souvik@ee.iitb.ac.in

***Abstract*— A cost-function based parameter optimizer is developed for the publicly available Compact-Physical (Comphy) framework and fit experimental Negative Bias Temperature Instability (NBTI) data. Ultrafast (10µs delay) measured threshold voltage shift (ΔV_T) during and after NBTI stress at various stress (V_{GSTR}) and recovery (V_{GREC}) bias, temperature (T) and stress time (t_{STR}), and mixed AC-DC stress using random V_{GSTR}, V_{GREC}, pulse duty cycle (PCD) and frequency (*f*) are used. Comparison is done with modeling using the BTI Analysis Tool (BAT) framework.**

Introduction: NBTI continues as a dominant reliability issue in modern p-FETs [1]. It causes positive charge buildup in the gate insulator of the device during stress, and shifts its parameters in time. The accrued charges partially reduce after the removal of stress, which results in recovery of parametric shift. Therefore, the modeling of NBTI becomes challenging, especially for gate pulses having arbitrary on (stress) and off (recovery) phases that mimic the data-paths in digital circuits, and for non-digital pulses encountered in analog and mixed-signal applications.

Although the physics of NBTI remains debated [2]-[4], there are two competing frameworks at present: BAT [1] and Comphy [5], the latter is publicly available in [6]. In BAT, uncorrelated contributions from interface trap generation (ΔV_{IT}), trapping of holes in pre-existing traps (ΔV_{HT}), and generation of bulk gate insulator traps (ΔV_{OT}) are used to model the ΔV_T time kinetics during and after stress. Reaction-Diffusion (RD) model together with Transient Trap Occupancy Model (TTOM) is used for ΔV_{IT}, Activated Barrier Double Well Thermionic (ABDWT) model for ΔV_{HT}, and Reaction-Diffusion-Drift (RDD) model for ΔV_{OT}, the model details and parameters are provided in various chapters of [1]. On the other hand, Comphy uses uncorrelated contributions from the recoverable (R) and (semi) permanent (P) components, handled respectively by Non-radiative Multi-Phonon (NMP) and Two Well Thermionic (TWT) models. In modern p-FETs having High-K Metal Gate (HKMG) stacks, R is individually calculated for the interlayer (IL) and High-K layer.

BAT is successfully validated by using diverse experimental conditions and across various technologies [1]. In this work, the Comphy framework is evaluated using basic DC stress-recovery and mixed DC-AC stress. The BAT results from [1] are shown as a comparative reference.

Modeling: Comphy has *19* adjustable parameters: *6×2* for R and *7* for P. A Root Mean Square Error (RMSE) cost function based optimizer is developed for fitting Comphy simulation with data, and all *19* parameters are freely varied. BAT has a maximum of *14* adjustable parameters, however, not all are freely varied, see [1] for details. For both models, once a parameter set is obtained for a particular device, they are kept fixed to simulate different experimental conditions.

Measured ΔV_T time kinetics (at fixed V_{GSTR}, T) and fixed time ΔV_T versus V_{GSTR} and T from Replacement Metal Gate (RMG) HKMG FinFETs are modeled respectively in Fig.1 and Fig.2 by BAT and Comphy, and the subcomponents are shown. For BAT, during stress, ΔV_{HT} saturates early, while ΔV_{IT} and ΔV_{OT} evolve in time, respectively with long-time power-law slope of ~1/6 and ~1/4; after stress, ΔV_{HT} and ΔV_{OT} respectively recover fast and show negligible recovery, while ΔV_{IT} recovers over an extended period; ΔV_{IT} dominates overall ΔV_T (unless at high V_{GSTR} and/or T). For Comphy, during stress, R evolves with a shallower time slope than P, while recovery is only due to R; equal contribution is made by R and P (unless at high T when P dominates). For the same device, measured and modeled ΔV_T time kinetics during and after stress at fixed V_{GSTR} and varying T (Fig.3), and at fixed T and varying V_{GSTR} (Fig.4), and fixed time ΔV_T versus V_{GSTR} at various T (Fig.5) are shown. The optimized parameters are used to model recovery kinetics at different t_{STR} (Fig.6), and different V_{GREC} (Fig.7).

Gate First (GF) HKMG data are also modeled (Fig.8-Fig.11). Two sets of (19) parameters are obtained for Comphy, by fitting only ΔV_T stress-recovery time kinetics (DC optimized, Fig.8) or also with DC multicycle data (Total optimized, Fig.9, Fig.10 (a), (b)). DC multicycle simulation (Fig.10 (a) and (b)) are shown for both sets of parameters. The Total optimized parameters are then used for simulation of mixed DC-AC in Fig.10 (c)-(f) and Fig.11 (a) and (b), and AC multicycle with different V_{GREC}, *f* and PDC in Fig.11 (c)-(f). For BAT, only 10 parameters are adjusted [1].

Conclusion: BAT and Comphy can model simple time kinetics during and after NBTI stress at different V_{GSTR} x T. However, Comphy faces some challenges in modeling recovery at different V_{GREC} (for analog applications), as well as arbitrary and mixed DC-AC gate waveforms (for digital data-path signals), although BAT can model the same. Comphy needs to be further tested by using higher *f* AC stress experiments (like BAT [1]) for further verifications. However, higher *f* AC simulation is difficult with publicly available Comphy due to large memory consumption.

References: [1] Mahapatra (Ed.), Springer, 2022, [2] Mahapatra, TED, 2013, p. 901, [3] Stathis, MR, 2018, p. 244, [4] Mahapatra, MR, 2018, p. 127, [5] Rzepa, MR, 2018, p. 49, [6] https://comphy.eu/downloads, [7] Parihar, TED, 2018, p. 23, [8] Parihar, TED, 2018, p. 392.

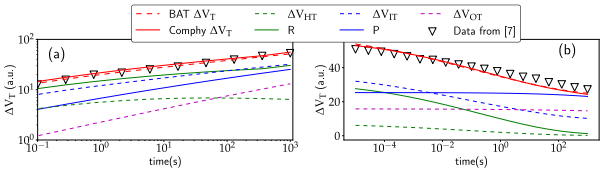


Fig.1 Measured (symbols) and modeled (lines) ΔV_T time evolution at fixed $V_{GSTR} = -1.7V$ and T=100°C during (a) stress and (b) subsequent recovery in RMG FinFET.

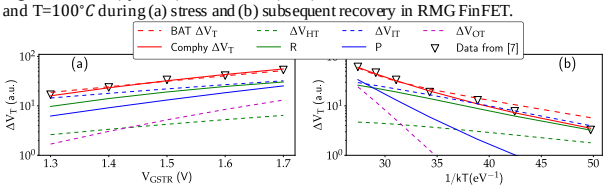


Fig.2 Measured (symbols) and modeled (lines) fixed time ΔV_T at t_{STR}=1Ks, as a function of (a) V_{GSTR} at T =100°C, (b) T at V_{GSTR} =−1.5V. in RMG FinFET.

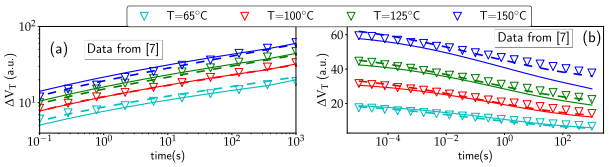


Fig.3 Measured (symbols) and modeled (lines) ΔV_T time evolution at fixed $V_{GSTR} = -1.5V$ for different T during (a) stress and (b) subsequent recovery in RMG FinFET. Solid and Dashed lines are for Comphy and BAT respectively.

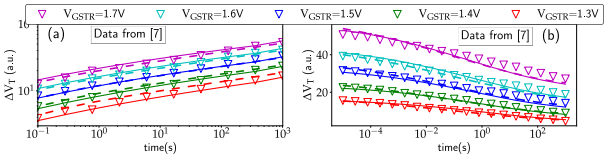


Fig.4 Measured (symbols) and modeled (lines) ΔV_T time evolution at fixed $T=100°C$ for different V_{GSTR} during (a) stress and (b) subsequent recovery in RMG FinFET devices. Solid and Dashed lines are for Comphy and BAT respectively.

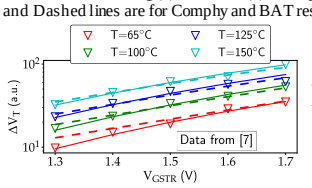


Fig.5 Measured (symbols) and modeled (lines) fixed time ΔV_T at t_{STR}=1Ks, at different T, as a function of V_{GSTR}, in RMG FinFET. Solid and Dashed lines are for Comphy and BAT respectively.

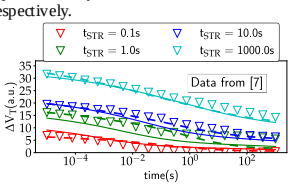


Fig.6 Measured (symbols) and modeled (lines) ΔV_T time evolution at fixed V_{GSTR} = −1.5V and T=100°C but different t_{STR} in RMG FinFET. Solid and Dashed lines are for Comphy and BAT respectively.

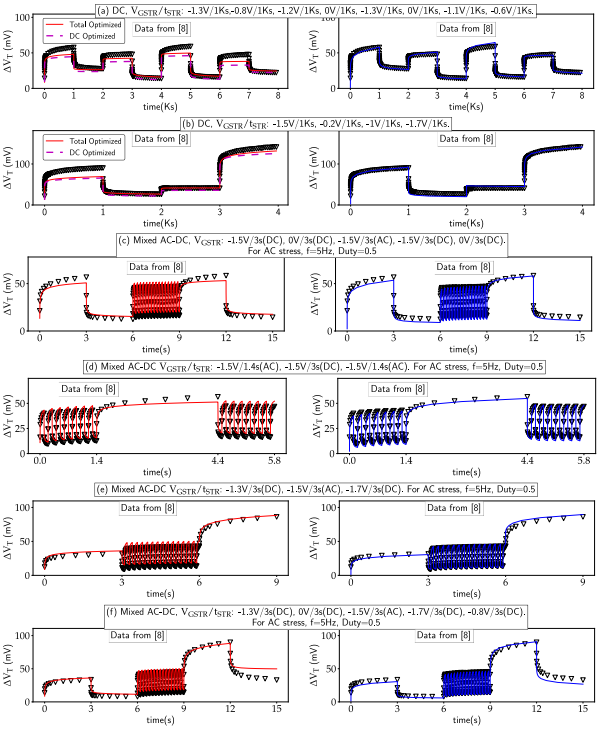


Fig.10 Measured (symbols) and modeled (lines) of arbitrary ΔV_T in GF planar MOSFET (a) & (b) multiple dc segments with different V_{GSTR} but fixed t_{STR}, (c) mixed ac–dc stress with inserted recovery after dc stress and fixed V_{GSTR} and t_{STR}, (d) mixed ac–dc stress with ac stress before dc stress and fixed V_{GSTR} and t_{STR}, (e) mixed ac-dc stress with ac between dc stress and varying V_{GSTR}, (f) mixed ac-dc stress with inserted recovery after dc stress and varying V_{GSTR}. Comphy (left panel) and BAT(right panel) are shown.

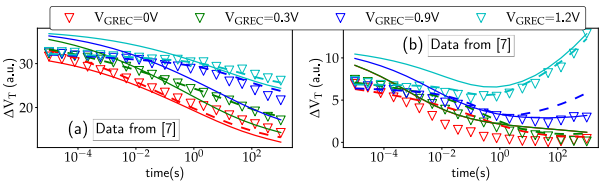


Fig.7 Measured (symbols) and modeled (lines) ΔV_T time evolution at fixed $V_{GSTR} = -1.5V$ and T=100°C but different V_{GREC} in GF planar MOSFET for (a) t_{STR} = 1000s, (b) t_{STR}=0.1s. Solid and Dashed lines are for Comphy and BAT respectively.

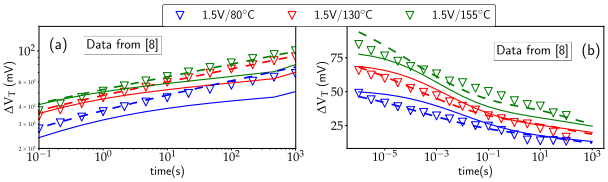


Fig.8 Measured (symbols) and modeled (lines) ΔV_T time evolution during (a) stress and (b) subsequent recovery in GF planar MOSFET, with DC optimized parameters for Comphy. Solid and Dashed lines are for Comphy and BAT respectively.

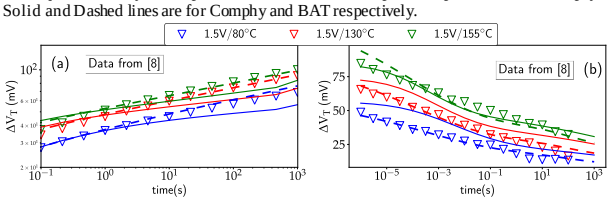


Fig.9 Measured (symbols) and modeled (lines) ΔV_T time evolution during (a) stress and (b) subsequent recovery in GF planar MOSFET, with Total optimized parameters for Comphy. Solid and Dashed lines are for Comphy and BAT respectively.

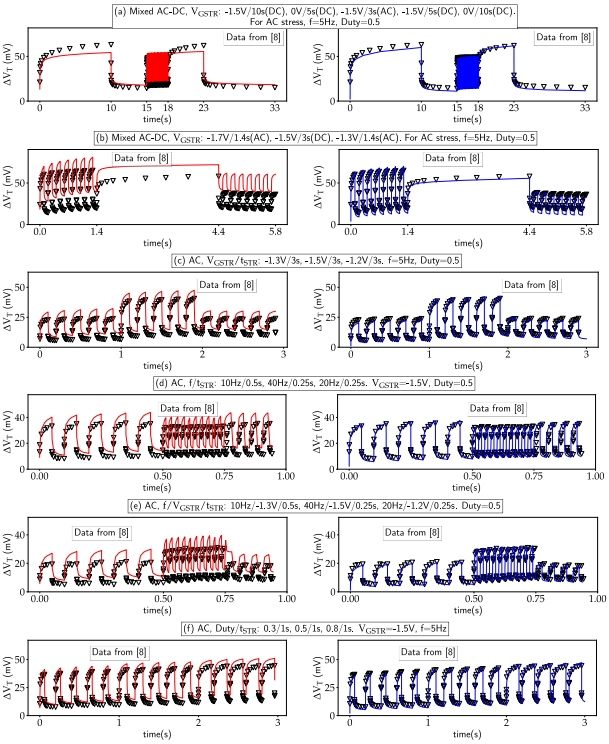


Fig.11 Measured (symbols) and modeled (lines) of arbitrary ΔV_T in GF planar MOSFET (a) mixed ac–dc stress with inserted recovery after dc stress with fixed V_{GSTR} and varying t_{STR}, (b) mixed ac–dc stress with ac stress before dc stress with varying V_{GSTR} and t_{STR} (c) multiple ac stress with varying V_{GSTR}, (d) multiple ac stress with varying frequency, (e) multiple ac stress with varying V_{GSTR} and frequency, (f) multiple ac stress with varying PDC. Comphy (left panel) and BAT(right panel) are shown.

Disorders in δ-layer tunnel junctions

Juan P. Mendez and Denis Mamaluy
Sandia National Laboratories
1515 Eubank SE, Albuquerque, NM 87123
email: jpmende@sandia.gov

Introduction

Phosphorus δ-layers in silicon (Si:P δ-layer) and, in particular, P δ-layer tunnel junctions (Si:P δ-layer tunnel junctions) have raised a lot of interest due to their high potential to become one of the important building blocks for quantum and beyond Moore computing applications.

Atomic precision advanced manufacturing (APAM) [1], which can be used to create doped planar structures (e.g. P or B) in Si with sub-nanometer abruptness precision and single-donor precision, provides a platform for manufacturing these types of systems. However, unintentional disorders can be introduced during the manufacturing process due to the stochasticity of the chemistry underlying the APAM process or due to the distinct thermal processes.

Previous theoretical approaches to investigate δ-layer systems [2,3] were based on the employment of periodic boundary conditions along the propagation direction, which become inapplicable in the case of tunnel junctions (TJs). Additionally, in our previous works [4], we have shown the need for an open-system quantum transport (QT) treatment to study highly conductive δ-layer systems.

In this work, we apply an efficient self-consistent QT formalism [5,6,7,8] to investigate how disorders might alter the tunneling current in Si: P δ-layer TJs.

Quantum Transport Framework

Our open-system QT framework [7,8,9,10,11] relies on a self-consistent solution of the Poisson-open system Schrödinger equation in the effective mass approximation and the Non-Equilibrium Green's Function (NEGF) formalism.

Within the NEGF formalism, the current $J_{\lambda\lambda'}$ from lead λ to λ' is computed from the Landauer formula

$$J_{\lambda\lambda'} = \frac{2e}{h}\int T_{\lambda\lambda'}(E)(f_\lambda(E) - f_{\lambda'}(E))dE, \qquad (1)$$

where e is the electron charge, h is the Planck's constant, E is the energy, $f_\lambda(E) = f(E - E_F - qV_\lambda)$ is the Fermi-Dirac function within the leads, V_λ is the applied voltage to the lead, E_F is the Fermi level and $T_{\lambda\lambda'}$ is the electronic transmission from λ to λ'. The transmission function is given by

$$T_{\lambda\lambda'}(E) = Tr(\Gamma_\lambda \mathbf{G}_D \Gamma_{\lambda'} \mathbf{G}_D^\dagger), \qquad (2)$$

where $\Gamma_\lambda = i(\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^\dagger)$ are the coupling ($N_D \times N_D$)-matrices between the device and the leads, and $\mathbf{G}_D$ and $\mathbf{G}_D^\dagger$ are the retarded and advanced Green's functions ($N_D \times N_D$)-matrices of the coupled device with the leads (open-system device). The retarded Green's function matrix can be computed using the Dyson equation

$$\mathbf{G}_D = [\mathbf{I} - \mathbf{G}_D^0 \mathbf{\Sigma}]^{-1} \mathbf{G}_D^0 \qquad (4)$$

where $\mathbf{G}_D^0 = \sum_\alpha \frac{|\Phi_\alpha\rangle\langle\Phi_\alpha|}{E^+ - E_\alpha}$ and $\mathbf{\Sigma} = \sum_{\lambda=\lambda_1}^{\lambda_L} \mathbf{\Sigma}_\lambda$. The electron density matrix is defined as

$$\rho(r_i) = \sum_\lambda \int_{-\infty}^{\infty} \rho_\lambda(r_i, E) f_\lambda(E) dE, \qquad (5)$$

where $\rho_\lambda(r_i, E) = \frac{1}{2\pi} \mathbf{G}_D \Gamma_\lambda \mathbf{G}_D^\dagger$. We can notice that all matrices in the above matrix operations are of size ($N_D \times N_D$), where N_D is the total grid-points of the discretized device domain. Thus, for instance, the inversion matrix cost in Eq. 4 is of $O(N_D^3)$, and the calculation cost of the eigenstates of $\mathbf{H}_D^0$ in Eq. 4 is of $O(N_D N_e^2)$, where N_e is the number of calculated eigenstates.

To reduce the computational cost of these intensive calculations, we utilize the Contact Block Reduction (CBR) method. The CBR is an efficient method to calculate the electronic transmission function of an arbitrarily shaped, multi-terminal open device. Within this method, the N_D grid-points are subdivided into N_C boundary grid-points with the leads and N_{Di} interior grid-points of the device domain ($N_D = N_C + N_{Di}$, $N_{Di} >> N_C$). With this domain discretization, the self-energy matrix $\mathbf{\Sigma}$, the open-system device Hamiltonian $\mathbf{H}_D$, and the Green's function matrix of the open-system device $\mathbf{G}_D$ can be expressed as submatrices

$$\mathbf{H}_D = \begin{pmatrix} \mathbf{H}_C & \mathbf{H}_{CDi} \\ \mathbf{H}_{DiC} & \mathbf{H}_{Di} \end{pmatrix}, \mathbf{G}_D = \begin{pmatrix} \mathbf{G}_C & \mathbf{G}_{CDi} \\ \mathbf{G}_{DiC} & \mathbf{G}_{Di} \end{pmatrix} \text{ and } \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_C & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where the size of the sub-matrices $\mathbf{H}_C$, $\mathbf{G}_C$, and $\mathbf{\Sigma}_C$ are ($N_C \times N_C$), the size of the sub-matrices $\mathbf{H}_{CDi}$, $\mathbf{G}_{CDi}$ and $\mathbf{\Sigma}_{CDi}$ are ($N_C \times N_{Di}$), and the size of the submatrices $\mathbf{H}_{Di}$, $\mathbf{G}_{Di}$ and $\mathbf{\Sigma}_{Di}$ are ($N_{Di} \times N_{Di}$). After some algebra, the electronic transmission from lead λ to λ' can be computed as

$$T_{\lambda\lambda'}(E) = Tr\left(\Gamma_{C_\lambda} \mathbf{G}_C \Gamma_{C_{\lambda'}} \mathbf{G}_C^\dagger\right), \qquad (6)$$

where $\mathbf{G}_C = [\mathbf{I} - \mathbf{G}_C^0 \mathbf{\Sigma}_C]^{-1} \mathbf{G}_C^0$ and $\Gamma_{C_\lambda} = i(\mathbf{\Sigma}_{C_\lambda} - \mathbf{\Sigma}_{C_\lambda}^\dagger)$. Similarly, the electron density matrix can be computed as

$$\rho(r_i) = \int_{-\infty}^{\infty} \Xi(E) f_\lambda(E) dE, \qquad (7)$$

where $\Xi(E) = \frac{1}{2\pi} \frac{\mathbf{B}_C^{-1} \Gamma_C \mathbf{B}_C^{-1\dagger}}{(E^+ - E_\alpha)(E^- - E_\alpha)}$ and $\mathbf{B}_C = \mathbf{1}_C - \mathbf{\Sigma}_C \mathbf{G}_C^0$. Importantly, note that all matrices are now of size ($N_C \times N_C$), where $N_{Di} >> N_C$. Therefore, we can reduce considerably the computational cost by employing the CBR method in the NEGF formalism.

For the self-consistent solution of the non-linear Poisson equation, we employed a combination of the predictor-corrector approach and the Anderson mixing scheme.

Results and discussion

We have applied our QT framework to investigate disorders in Si: P δ-layer TJs. The device is shown in Fig. 1. The considered disorders include variations of the δ-layer thickness (the ideal δ-layer thickness corresponds to the mono-atomic layer, which is approximately 0.2 nm), small variations of the tunnel gap lengths ($L_{gap} + \delta_{Lgap}$) from the desired lengths, and the presence of impurities and

charged dipoles in the tunnel gap. To carry out this study, we have evaluated the tunneling current ratio between the "non-ideal" device (with disorder) and the "ideal/reference" device (without disorder) for a wide range of tunnel gap lengths.

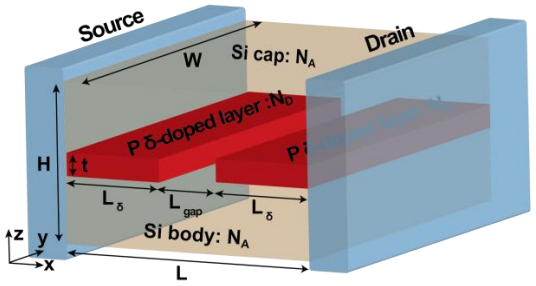
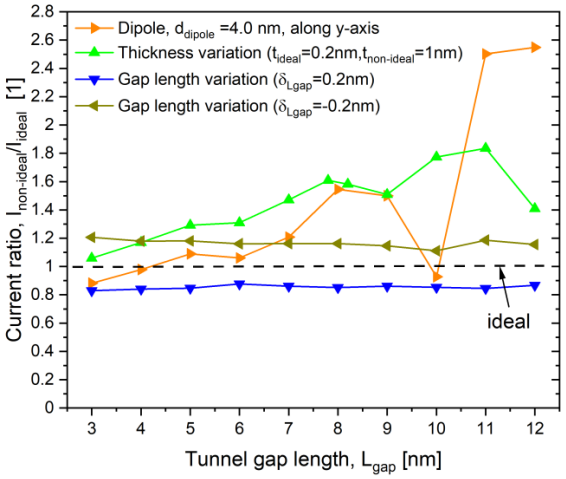


Fig. 1: The Si:P δ-layer TJ device is composed of a Si body, a very high P-doped layer with an intrinsic gap, and a Si cap.

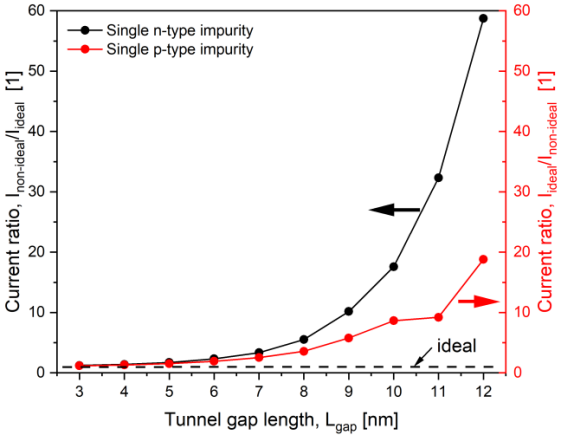
Fig. 2 shows the tunneling current ratio for the distinct disorders mentioned above. Firstly, one can observe that small deviations from the ideal δ-layer thickness (see green curve in Fig. 2(a)) and tunnel gap lengths (see olive and blue curves in Fig. 2(a)) can only alter the tunneling rate up to 20%. On contrary, the presence of impurities near the intrinsic gap strongly alters the tunneling rate (see red an black curves in Fig. 2(b)), especially for large tunnel gaps (L_{gap}>10nm). Our simulations also predict a strong asymmetry with the impurity electrical sign: n-type impurities in the intrinsic tunnel gap dramatically affect the current value, increasing the current; at the same time, p-type impurities reduce the current to a significantly smaller manner. In Fig. 2(a), it also includes the effect of dipoles in the tunnel gap, i.e. when two distinct electrical sign-charged impurities are in proximity. As expected the effect of a dipole is significantly weaker than that of a single charged impurity: the strong increase in the tunneling rate by the presence of an n-type impurity is neutralized by the decrease in the tunneling rate by the presence of a p-type, especially for short dipole lengths. For very large dipole lengths, where there is no interaction between the positive and negative charges, the effect on the tunneling rate is a superposition of both, thus n-type impurity taking prevalence. It is also evident that the effect of a dipole in the tunnel gap strongly depends on the dipole moment orientation. Overall, however, the effect of a dipole in the gap on the tunneling rate is weaker than the effect of a single charged impurity in the tunnel gap.

Conclusions

We have employed an efficient self-consistent implementation of the NEGF formalism. We have applied this framework to investigate the effect of disorders on the tunneling current in δ-layer TJs. We have considered variations of the δ-layer thickness, small deviations of the tunnel gap lengths, and the presence of impurities and charged dipoles in the intrinsic tunnel gap. While most disorders moderately affect the tunneling rate (up to 20%), a single charged impurity in the tunnel gap can alter the tunneling rate by more than an order of magnitude, especially for relatively large tunnel gaps (>10nm). The electric sign of impurity plays an important role in the tunneling rate: the change of current due to an n-type impurity is stronger than for p-type impurity, especially for large tunnel gap lengths (>10 nm). Thus, we can conclude that our simulations suggest overall that geometric fidelity of the device fabrication is less important than mitigation of impurities inside of the junction. Finally, our simulations also shed light on some manufacturing challenges and guidance for the next generation of quantum and beyond-Moore devices using δ-layer as a building block.



(a)



(b)

Fig. 2: Effect of non-idealities on the tunneling current for an applied bias of 100 mV: (a) single n-type impurity and p-type impurity; (b) . $N_D = 10^{14}$ cm^{-2}, $N_A = 10^{17}$ cm^{-3}, t = 1 nm, L = 50 nm and H = 20 nm.

References

[1] D. Ward, *et al.*, EDFAAO 22, 4 (2020)
[2] S. Lee *at al*., Phys. Rev. B 84, 205309 (2011).
[3] D.J. Carter *et al.* Phys. Rev. B 79, 033204 (2009).
[4] D. Mamaluy, J. P. Mendez *et al.*,Commun Phys **4,**205(2021)
[5] D. Mamaluy *et al.*, J. Appl. Phys. 93, 4628–4633 (2003).
[6] D. Mamaluy *et al.*, Phys. Rev. B 71, 245321 (2005).
[7] H. R. Khan, *et al.*, IEEE T. Electron Dev. 54,784–796 (2007).
[8] D. Mamaluy, X. Gao, APL 106, 193503 (2015).
[9] X. Gao *et al.*, J. Appl. Phys. 115, 13707 (2014).

Ferroelectric FDSOI FET Modeling for Memory and Logic Applications

Swetaki Chatterjee[1], Shubham Kumar[2], Amol Gaidhane[3], Chetan Kumar Dabhi[4],
Yogesh S. Chauhan[2], Hussam Amrouch[1]
[2]Indian Institute of Technology, Kanpur; [3]Arizona State University; [4]University of California, Berkeley
[1]Chair of Semiconductor Test and Reliability, University of Stuttgart, Germany. Email: chattesi@iti.uni-stutgart.de

Abstract: In this paper, we present a Verilog-A based compact model for simulating Ferroelectric Fully Depleted Silicon-on-Insulator (Fe-FDSOI) FET. The model can capture the rich physics of ferroelectric (FE) materials and reproduce the important electrical characteristics, such as history effect, the impact of threshold voltage on pulse width and amplitude as well as potentiation-depression characteristics. The FE switching is modeled using Preisach model to capture the Polarization (P) – Voltage (V) characteristics. In addition, we capture the history-dependent minor loop characteristics to obtain multiple states of polarization. This allows the modeling of multiple conductance states, which form the fundamental prerequisite for neuromorphic applications as well as multi-level non-volatile memories. The underlying baseline FDSOI FET is modeled using the industry-standard BSIM-IMG compact model. The model is then augmented with the physics-based model of FE capacitor to realize Fe-FDSOI FET. Our model is computationally efficient and carefully calibrated to reproduce experimental measurement data.

Introduction: Fe-FDSOI shows a great potential for non-volatile memory and neuromorphic applications [1]. To fully exploit the functionality of FeFETs, we need efficient compact models to explore the large available design space. However, existing models require a significant amount of memory and time, which is computationally inefficient for large circuit simulations [2]. Hence, a highly efficient compact model is a key requirement to successfully reproduce the trend of memory window, conductance, and history effects from experimentally measured data. In this paper, we present an efficient compact model that stores a smaller number of turning points and can still reproduce all the experimentally observed characteristics. The primary advantage of our model comes from using R-C element to calculate the delay and store lesser number of turning points [3]. To validate the working of our model, we first calibrate the industry-standard compact model for the FDSOI technology (BSIM-IMG) with experimental data to calibrate the underlying FET device. Then, we calibrate the FE capacitor using the developed Fe-Cap model to get the required parameters for the realization and simulation of Fe-FDSOI. We then simulate the transfer characteristics of the device and the effect of back bias. Further, we show the history effects by showing pulse amplitude and width modulation. We finally demonstrate the potentiation and depression characteristics exhibited by the FDSOI which can be used for neuromorphic and other memory applications.

Results and Discussions: The structure of the Fe-FDSOI is shown in Fig. 1(a) in which MFMIS configuration is employed. As compared to the MFIS structure, this is computationally more efficient as it solves two different circuit entities simultaneously without sacrificing any important characteristics [3]. The underlying FDSOI FET is modeled using the industry-standard BSIM-IMG [4] compact model. It is calibrated against experimental data from [5], shown in Fig. 1(b). The n-FDSOI FET characteristics are obtained by mirroring the p-FDSOI FET characteristics. The Fe-Cap is modelled using Preisach model. The slope of the P-V curve (m) is calculated by considering the prior turning points. The turning points are calculated based on a delayed voltage (V_{del}) using an R-C delay network [3]. If V_{del} is greater than or less than the input voltage, we correspondingly calculate as turning up or turning down. Our Preisach model of the Fe-Cap is also validated against experimental data from [2] to get the required values of Remnant Polarization (P_r), Coercive Field (E_c) and Saturation Polarization (P_s), as Fig. 2 shows. We also demonstrate the minor loops by storing the turning points and calculating the slope of the loop based on stored values. However, compared to prior implementations that required storing of all the turning points, we store only the last two turning points. For cases when the voltage exceeds the stored turning point, we use the maximum positive or negative bias to calculate m. Hence, we calculate m and P at every bias point by:

$$m = \frac{P_{aux_u} - P_{aux_l}}{P_s\left(tanh\left(w(V_{aux_u} - V_c)\right) - tanh\left(V_{aux_l} - V_c\right)\right)} \quad (1)$$

$$P_{aux} = m * P_s \times tanh\left(w(V_{aux} - V_c)\right) + P_{off} \quad (2)$$

Where, u, l corresponds to the stored points and V_{aux} is the input voltage across the ferroelectric after relaxation. We then combine the model of the Fe-Cap and BSIM-IMG by solving the charge balance at every instance using commercial SPICE simulator to get Fe-FDSOI. The simulated Fe-FDSOI characteristics are shown in Fig. 3 with a memory window of about 1.2 V (when programmed with a pulse of magnitude = 5 V) between the high V_t and low V_t states. One of the important advantages of using Fe-FDSOI is the ability to control V_t using back-bias. The effect of back bias on the transfer characteristics is shown in Fig. 4. A negative back bias shifts the threshold voltage of both the states to the right, whereas a positive back bias shifts it towards the left. This is especially useful for some memory applications such as ternary content addressable memory to ensure reliable operation [6]. On applying input pulses of increasing magnitude sets the Fe-FDSOI in different states. Positive program pulse (Vp) and negative erase pulse (Vn) as shown in Fig. 5 (a) and (b) are applied, followed by a reading sweep. The results are shown in Fig. 5 (c) and (d), which clearly demonstrate different V_t states. Recently, Fe-FET has been demonstrated as an analog synapse for neuromorphic applications where potentiation and depression of the conductance states is an important step in weight update [1, 2]. We calculate the conductance at Vds= 0.1 V after every pulse, as shown in Fig. 6. It shows relatively linear and symmetric characteristics, which is highly desirable. Finally, we show the effect of pulse width modulation on the Fe-FDSOI in Fig. 7. Not much variation of the memory window is observed after the device switches. This reaffirms amplitude modulation is a more suitable scheme for updating weights in neuromorphic computing.

Conclusion: We presented a robust and efficient compact model for Ferroelectric FDSOI simulations. It paves the way for further development of large-scale circuits, including non-volatile memory, logic- and in-memory applications.

References: [1] H.Mulaosmanovic et al., IEEE VLSI Sym, 2017; [2] K.Ni et al., IEEE VLSI Symposium 2018 ; [3] A Gaidhane et al., TechRxiv, 2022; [4] H.Agarwal et al., IEEE EDTM 2020; [5] Q. Liu et al., IEE IEDM 2013; [6] X.Yin et al., DATE 2017.

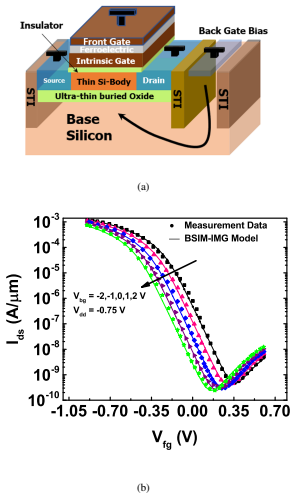


Fig. 1. (a) The simulated Fe-FDSOI structure with MFMIS configuration. (b) Calibration of Drain current (I_{ds}) - Front Gate voltage (V_{fg}) characteristics of pFET varying back gate bias Voltage (V_{bg}) from -2 V to 2 V with V_{dd} = 0.75 V with a gate length (L_g) of 20 nm. Measurement data is taken from [5].

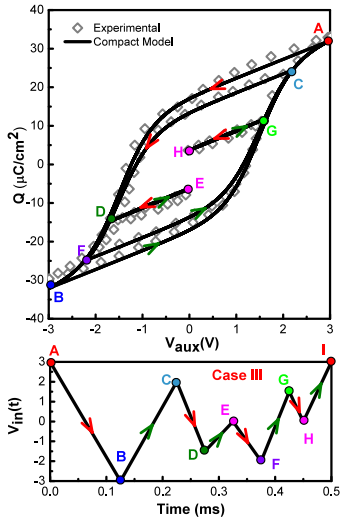


Fig. 2. Validation of simulated P_{aux}-V_{aux} with a 10nm HfO2 based ferroelectric material experimental results [2] for a non periodic signal to determine the turning points.

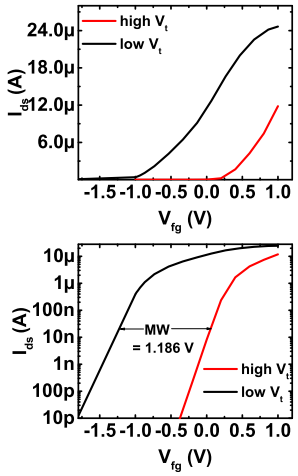


Fig. 3. The simulated transfer characteristics of the Fe-FDSOI device having a width of 100 nm in (a) Linear scale, and (b) Log scale showing a memory window (MW) of 1.186 V.

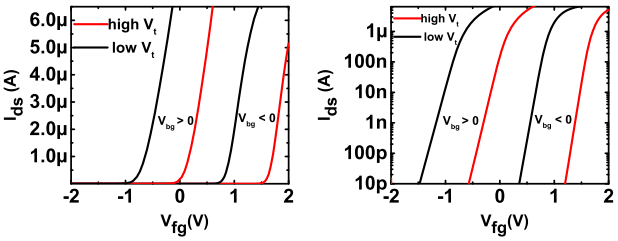


Fig. 4. Effect of back bias on the transfer characteristics of the Fe-FDSOI structure shown in (a) Linear scale (b) Log scale. A positive back bias shifts the graph to the left thus reduces V_{th} and vice-versa for negative back bias.

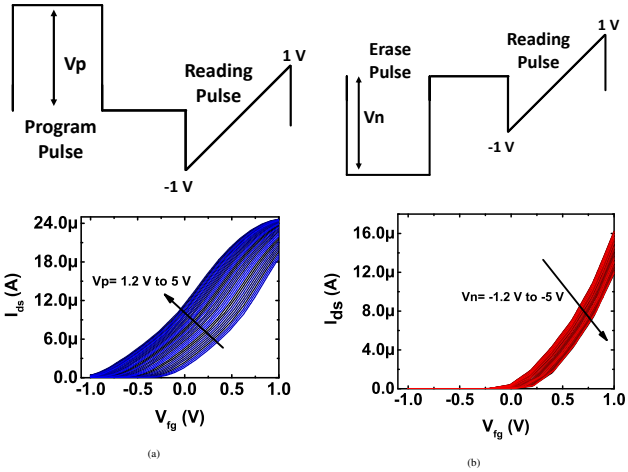


Fig. 5. The applied input pulses of varying positive amplitude Vp and negative amplitude Vn to set in different V_t for (a) programming and (b) erase states. The values of Vp and Vn set the device into different threshold voltage levels thus enabling MLC as well as for neuromorphic applications.

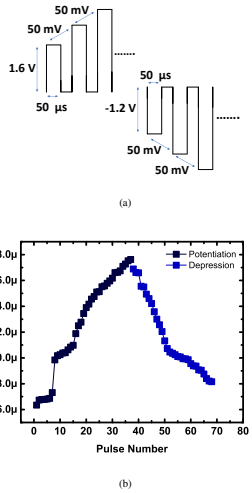


Fig. 6. (a) Applied waveforms for the channel conductivity increase (potentiation) and decrease (depression). (b) potentiation and depression characteristics. The conductance can be mapped directly to neural weights and can be used for neuromorphic applications.

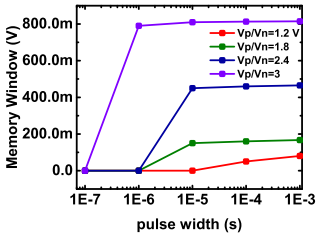


Fig. 7. Variation of memory window with applied pulse width for different pulse amplitude. After the device switches, the variation is very less.

Hierarchical Simulation of Nanosheet Field Effect Transistor: NESS Flow

D. Nagy*† A. Rezaei* N. Xeni* T. Dutta* F. Adamu-Lema*† I. Topaloglu* V. P. Georgiev* A. Asenov*†
*Device Modelling Group, School of Engineering, University of Glasgow, Glasgow, Scotland, UK
†Semiwise Ltd., Rankine Building, Glasgow, Scotland, UK
daniel.nagy@glasgow.ac.uk

I. Introduction

Nanosheet gate-all-around transistor devices have been adopted by Samsung in their 3 nm CMOS offering [1]. Compared to FinFETs they have superior electrostatic control [2], [3]. The nanosheet architecture can also be vertically stacked thus achieving higher drive current on a same footprint area compared to a single nanowire or nanosheet. Accurate device simulations are crucial for the development and the optimization of the nanosheet transistors. With this in mind, we have developed and report a hierarchical simulations flow implemented in the Glasgow Nano-Electronic Simulation Software (NESS) in order to enable the accurate simulation and optimization of the nanosheet transistors [4], [5].

II. Method

In Fig. 1 we have shown the modular structure of NESS. The focus in the current abstract is the flow of the essential modules for the accurate simulation of the nanosheet transistors. The relevant NESS modules include the structure generator (SG), Effective Mass Extractor (EME), Non-Equilibrium Green's Function (NEGF) and Drift-Diffusion (DD) module. All transport solvers work in a unified solution domain created by SG and are linked to a common Poisson Solver as shown in Fig. 1. In this section we briefly describe each module in the contest of the nanosheet transistors simulation. A very important step in the quantum transport simulation of semiconductor devices working in strong confinement regime is the extraction of the effective masses for specific device cross section. NESS includes the EME module which allows the calculation of the effective masses, based on results from first principle band-structure simulation (Fig 3). The EME module extracts the parabolic transport masses using parabola fitting to the E-k relations for each sub-band and the confinement masses from the differences between the lowest sub-band energies. The NEGF transport formalism provides a quantum treatment of electron transport to capture quantum phenomena such as tunneling, coherence, and particle-particle interactions in mesoscopic and nano-scale devices. We can simulate ballistic transport or can enable electron-phonon (e–ph) interactions within the self-consistent Born approximation by including the optical (Opt), acoustic (Ac), or optical-acoustic (Opt+Ac) e-ph scattering to study the transport in diffusive limit. The carrier density, potential profile, and the current are obtained by performing a self-consistent solution of the Poisson equation and the NEGF transport solver in coupled mode-space representation. The DD module solves the carrier continuity equations self-consistently with the Poisson's equation. It includes different mobility models to account for the impact of doping, and parallel and perpendicular electric fields.

III. Results

SG – The structure generated for this paper is shown in Fig. 2. The device has a source/drain and gate length of 10 nm, a channel width and height of 12 and 3 nm, respectively. The gate oxide thickness is 1 nm SiO_2. For the current study we used rectangular shape although more realistic cross sections can be handled in NESS. Fig 3, shows the comparison of the tight binding band-structure obtained from the QuantumATK [6] and the extraction from NESS [7], resulting to a good agreement between the minima and curvature of the band-structure. NEGF – The I_D-V_G characteristic at drain bias of 0.05 V, and 0.7 V are represented in Figs. 4, and 5, respectively. The effect of scattering is also demonstrated in the figures. In Figs. 6, and 7 depict the OFF- and ON-state local density of states (LDOS), average potential, and energy sub-band structure for low and high drain biases. DD – As shown in Fig. 8, the DD simulator can be calibrated to fit the NEGF results. This enables the use of DD simulator for situations where large number of simulations are needed e.g. in variability simulations where schemes like NEGF are computationally too expensive.

IV. Conclusion

In the current work we have successfully demonstrated the capabilities of NESS in the hierarchical predictive simulations of nanosheet transistors. We have carried out device simulations and showed that the more accurate NEGF simulations can be used for the calibration of the classical DD simulations within one single toolbox. Additionally we showed that the EME module can be used to extract the effective masses for confined structure like the nanosheet.

References

[1] G. Bae *et al.*, in *IEDM*, 2018, pp. 28.7.1–28.7.4.
[2] F. M. Bufler *et al.*, *IEEE TED*, vol. **67**, no. 11, pp. 4701–4704, 2020.
[3] D. Nagy *et al.*, *IEEE Access*, vol. **8**, pp. 53 196–53 202, 2020.
[4] C. Medina-Bailon *et al.*, *Micromachines*, vol. **12**, no. 6, 2021.
[5] S. Berrada *et al.*, *J Comput Electron*, vol. **19**, no. 9, p. 1031–1046, 2020.
[6] Synopsys. (2020) *QuantumATK version R-2020.09*. [Online]. Available: https://www.synopsys.com/silicon/quantumatk/
[7] O. Badami *et al.*, *Applied Sciences*, vol. **9**, no. 9, 2019.

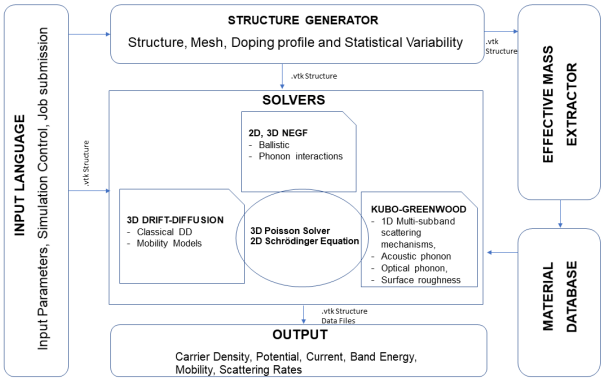


Fig. 1. NESS modular system.

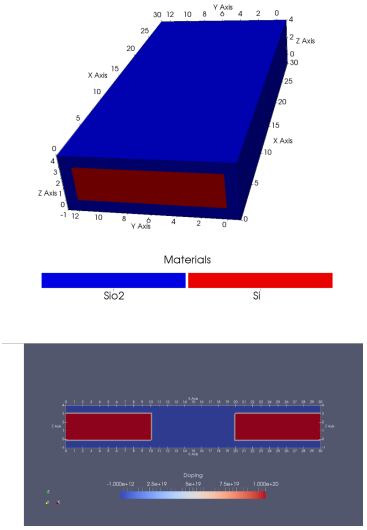


Fig. 2. Structure of nanosheet and doping profile along the channel.

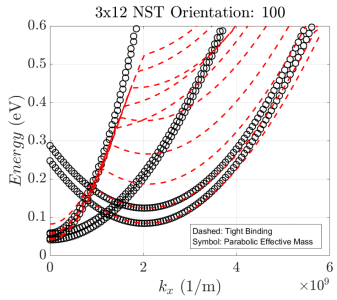


Fig. 3. Comparison of the electronic band-structure calculated using tight binding and the Parabolic Effective Mass approximation using the EME module for a 3nm×12nm Nanosheet with orientation along 100 direction.

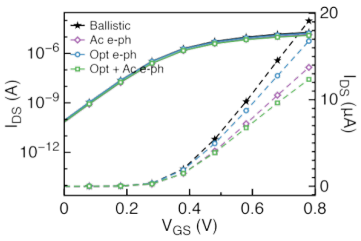


Fig. 4. I_D-V_G characteristics for ballistic and diffusive NEGF simulations at low drain (0.05V) bias. The both linear and logarithmic scale have been depicted.

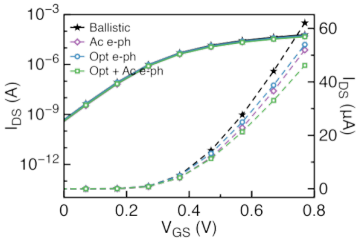


Fig. 5. I_D-V_G characteristics for ballistic and diffusive NEGF simulations at high drain (0.7V) bias. The both linear and logarithmic scale have been depicted.

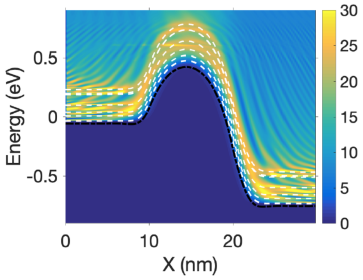


Fig. 6. The LDOS, average potential (black line), and energy sub-bands structure in the OFF-state in the ballistic limit for $V_D = 0.7$ V and $V_G = 0$ V.

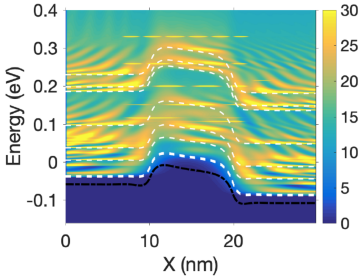


Fig. 7. The LDOS, average potential, and energy sub-bands structure in the ON-state with Opt+Ac e-ph scattering for $V_D = 0.7$ V and $V_G = 0.7$ V. The reference in energy is taken at the Source Fermi level ($E_{FS} = 0$ eV).

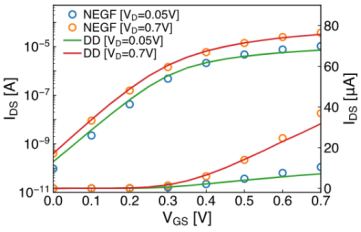


Fig. 8. Calibration of drift-diffusion simulator to NEGF results (with optical and acoustic phonon scattering) for I_D-V_G characteristics obtained at low and high drain bias.

Improvement of On-cell Metrology Using Spectral Imaging with TCAD Modeling

Byungseong Ahn[1], Kwangseok Lee[1], Jaehun Yang[1], Jiseong Doh[1], Jaehoon Jeong[1], Minseok Kim[2], Yeonjeong Kim[2], Jongchul Kim[2], Hyung Keun Yoo[2], Dae Sin Kim[1]

[1]Computational Science and Engineering Team, [2]Foundry Metrology & Inspection Team / Device Solution Business
Samsung Electronics Co., Ltd., Gyeonggi-do 18448, Republic of Korea.
* Email Address: bseong.ahn@samsung.com (B. Ahn).

Abstract

On-cell metrology based on spectral imaging (SI) with TCAD modeling is developed for measuring advanced logic devices. Spectral imaging has been adopted because it can directly measure a cell-block of logic device, which was difficult with the conventional methodology. One of conventional methodology, optical critical dimension (OCD), assumes that the cell-block is infinitely periodic, so that metrological site is needed and relatively large. SI has also disadvantages in consistency because it only measures a single spectrum type and narrow wavelength band. To overcome such limitations, we adopt TCAD modeling by providing realizable spectrum and structure. TCAD is modeling technique which creates a structure by physical or mathematical simulation. After generating pairs of spectrum and structural parameter, a correlation is obtained through machine learning. With proposed method, we could obtain both wafer-to-wafer distribution and in-wafer distribution with a good tendency in logic devices.

Keywords — Spectral Imaging, Optical critical dimension, Spectroscopic ellipsometry, Spectroscopic reflectometry

1. Introduction

Non-destructive in-line metrologies of logic devices are becoming important to prevent yield loss and evaluate electric characteristics. As the device scaling decreases and the development cycle shortens, its importance and demand is increasing further [1-2]. Due to this demand, measurement at the cell level is required to obtain the local characteristics which could not be achieved with optical conventional methodology. One of conventional methodology, optical critical dimension (OCD) [3-4] needs a large measurement area, assuming that the cell-block is infinitely periodic. Owing to its assumption and large area, the homogenized characteristics can be obtained, but the local characteristics of the cell-level cannot be achieved [5].

To overcome such limitations, we propose the spectral imaging (SI) with TCAD modeling. SI system is on-cell metrology which can generate high resolution spatial images. From analyzing spatial images, it can obtain local characteristics of cell-block. SI system measures only the reflectance of s-waves, sR, for the fast capturing, so that it has a disadvantage in terms of machine learning. This lack of spectrum type can be overcome through TCAD modeling by realizing structure and spectrum of simulation. The relationship between spectrum and structural parameter has been improved, and the wafer distribution tendency has been improved also.

2. Theoretical Backgrounds

2.1. Optical System Configuration

The schematic OCD and SI system is depicted in Fig. 1. As shown in Fig. 1(a), OCD system only one large broadband point source to wafer. The reflected light is split through the prism. The line charged-coupled device (CCD) received broadband spectrum. On the other hand, SI system incident single wavelength but multi-points source on wafer by the monochromater in Fig. 1(b). The reflected light is captured as an image by the area CCD. The stack of images stored by wavelength becomes a spectral cube as shown in Fig. 2(a). When gather information according to the wavelength of the spectral cube at an arbitrary position, each spectrum can be obtained (See Fig. 2(b)).

2.2. Methodology of SI system

SI system can acquire spectrum at high resolution multi-points, but it has a relatively lack of data from its configuration. The wavelength band of SI has a relatively coarse and narrow. Also, SI generally uses one type of spectrum type (reflectance of s-wave, sR) owing to the burden of data. Therefore, the learning process is necessary to complement the lack of data.

In the conventional SI approach, sR signal and the module target spec (MTS) are learned (See Fig. 3(a)). It does not require modeling, so machine learning is performed directly. MTS can be obtained from the transmission electron microscope (TEM) and scanning electron microscope (SEM), but it is hard to collect sufficient data for learning.

The key difference of proposed method is adopting TCAD model as shown in Fig. 3(b). TCAD model is a time-transient physical model that is possible to implement incoming factors. TCAD model is completed through three steps. In structural calibration, structure is generated suitable for TEM. The simulation spectrum can be calculated from the rigorous coupled-wave analysis (RCWA) and compared to the actual spectrum to define the good of fitness (GOF). In OCD signal GOF optimization, the spectroscopic ellipsometry (SE) and spectroscopic reflectometry (SR) are fitted. In SI signal GOF optimization, only sR signal is optimized with narrowing the wavelength band. TCAD model is free to construct simulation library which is pairs of spectrum and MTS. Fully connected neural networks (FCNNs) is adopted for the machine learning.

3. Result

The proposed method was used in a specific step of logic products. Fig. 4 shows the results of OCD and SI GOF

optimization. They satisfies GOFs of 0.98 or more respectively. Machine learning results have accuracy of R^2=0.53. Fig. 5 shows the results of wafer distribution according to etch time split. Wafer-to-wafer distribution, which is MTS change between other wafers is clearly distinguished. As the etching time increases, the CD becomes smaller. In-wafer distribution which is MTS change in wafer is confirmed clearly. The MTS change according to the radius is the same as the left and right, and it tends to be similar in other wafers.

4. Conclusion

In this paper, we propose a spectrally resolved imaging system with TCAD modeling. Through TCAD model, it can suggest better learning data similar to actual spectrum and MTS. From improved consistency and locality, proposed method can grasp early structural defects and contribute to yield improvement.

5. Reference

[1] Ha, D., et al. *VLSI Technology.* IEEE (2017)
[2] Singh, N., et al. *IEEE Electron Device Letters* (2006): 383-386.
[3] Zhou, R., et al. *International Society for Optics and Photonics* (2015): 942416
[4] Yazaki, A., et al. *Applied Physics Letters* (2014): 251106.
[5] Hodges, J. Scott, et al. *Metrology, Inspection, and Process Control for Microlithography XVII* (2003): 215-223

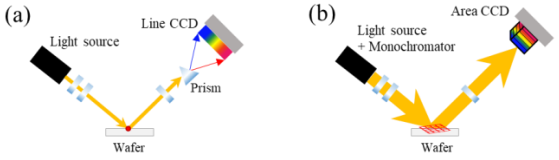


Fig. 1. Comparison between OCD system and spectral imaging system.

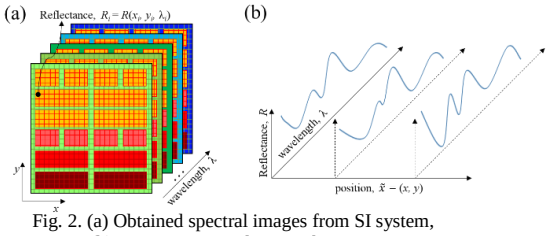


Fig. 2. (a) Obtained spectral images from SI system, (b) Spectra extracted in an arbitrary position

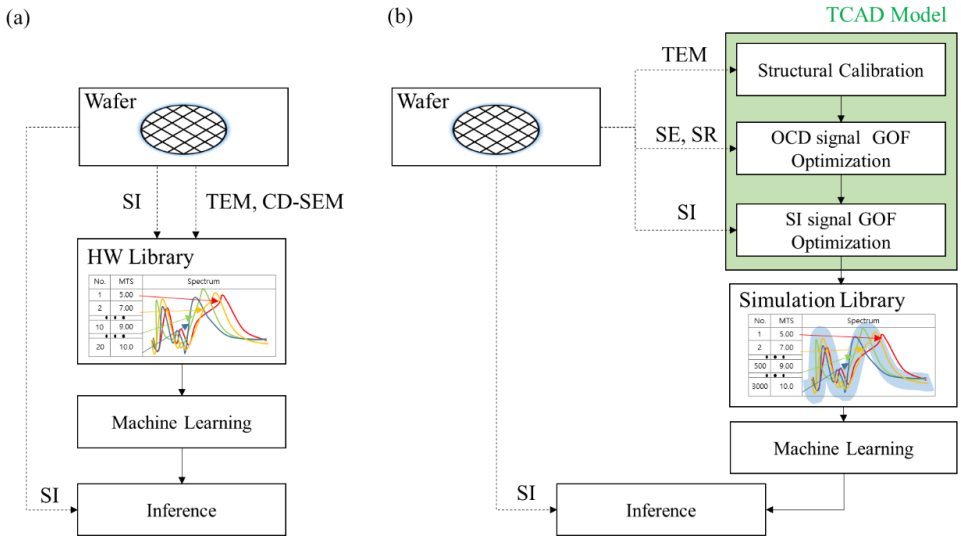


Fig. 3. Process flow of (a) conventional SI approach and (b) proposed method.

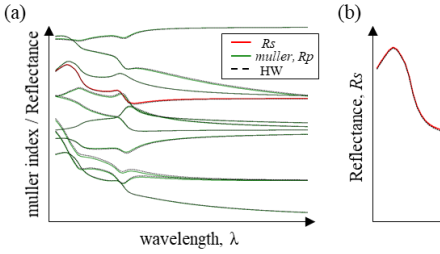


Fig. 4. GOF optimized results of (a) OCD case and (b) SI case.

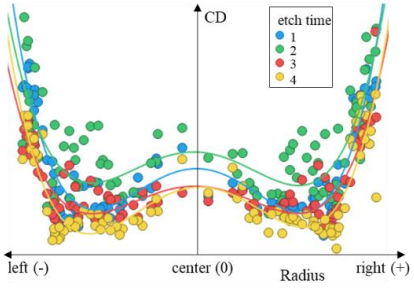


Fig. 5. Wafer distribution according to etch time

Modeling Optical Second Harmonic Generation for Oxide Semiconductor Interface Characterization

Binit Mallick[1], Dipankar Saha[1], Anindya Dutta[2], Swaroop Ganguly[1*]
[1]Department of Electrical Engineering, Indian Institute of Technology, Bombay, India
[2]Department of Chemistry, Indian Institute of Technology, Bombay, India
swaroop.ganguly@gmail.com

Contactless and non-invasive characterization of insulator-semiconductor interface can facilitate the development of a range of important semiconductor devices. Optical second-harmonic generation (SHG) is a sensitive technique well-suited for characterizing such layered structures. The dependency of SHG on the doping concentration in silicon (Si) [1], silicon dioxide (SiO_2) thickness [2], and measuring conduction band offset at the Si/SiO_2 interface [2] has demonstrated its potential and advantages over traditional capacitance/conductance as well as X-ray Photoemission Spectroscopy (XPS) and Internal Photoemission (IPE) methods for in-situ characterization. Mapping of carrier dynamics [3] and trapped charge [4][5][6] have also been demonstrated using this technique. While SHG based semiconductor/dielectric interface, specifically Si/SiO_2, characterization has been under development [7][8][9][10][11], the method is still not standardized enough to allow easy extraction of quantitative interface characteristics. This is due in part to the absence of models that are handy yet comprehensive. The pioneering effort by Mihaychuk *et al*. [7], for example, does not include interface charge densities, which are seen to be important in our experimental samples. In this work we present a comprehensive numerical model for the time-dependent second harmonic photon count obtained from our in-house experimental setup, which is being reported separately [12]. In what follows, the symbols may be understood to have their usual meanings.

The SHG process involves doubling the frequency of incident light, as a nonlinear medium converts the cumulative energy of two photons of the incident beam into a single photon [13][14]. In silicon, the nonlinearity arises due to the broken symmetry at the surface and an interface with other materials (here SiO_2 [7][8]). The measured second harmonic photon count profile is presented in Fig. 1a, along with a schematic in its inset. Its time-dependent nature arises from the trapping of photo-excited electrons at the interface, within the oxide region, and at the SiO_2 surface – illustrated in Fig. 1b. The trapping of these photo-excited electrons over time leads to the generation of a time-dependent electric field $\xi_{0-}(t)$ at the Si/SiO_2 interface. As the electric field $\xi_{0-}(t)$ rises, it contributes to the electric field-induced second harmonic generation (EFISHG) and the gradual increase in second harmonic intensity over time observed here.

A semi-empirical equation of the form $I^{(2\omega)}(t) = A_0 + A_1 e^{\frac{-t}{\tau_1}} + A_2 e^{\frac{-t}{\tau_2}} + A_3 e^{\frac{-t}{\tau_3}}$ is first used to fit the experimental time-dependent SH intensity data, where the parameters τ_1, τ_2, and τ_3 are understood to represent the trapping time constants at the Si/SiO_2 interface, within the oxide, and at the surface of SiO_2 respectively $\left(Q_{it}^{Trapped}(t), Q_{ox}^{Trapped}(t), Q_{st}^{Trapped}(t)\right)$. Their values signify the time taken to fill a portion (63%) of the available trap states at various positions mentioned above. The time constants are fed into the a numerical solver to calculate the time-dependent trapped charge and electric field $\xi_{0-}(t)$.

We have developed an algorithm to solve the Poisson-Boltzmann equation using the Newton-Raphson method at different time instances with the aforementioned trapped charges at that time instant. The flowchart thereof is shown in Fig. 2. The parameters used will be presented. In terms of validation by benchmarking to existing TCAD software, the available possibility was a steady-state comparison with predefined interface and oxide charges – this is shown in Fig. 3. The interface charge $Q_{it}\{\phi\}$ and potential profile $\phi\{Q_{it}\}$ in our model are intertwined in a self-consistent manner to capture the occupation probability of the interface states at different time instances. We have considered an interface trap state distribution similar to Thoan *et al*. [15] because of the similarity of their fabrication process to ours. The Poisson-Boltzmann solver yields the time evolution of the electric field $\xi_{0-}(t)$ and eventually the EFISHG intensity using $I^{(2\omega)}(t) = \eta|\chi^{(2\omega)} + \chi^{(3\omega)}\xi_{0-}(t)|^2\left(I^{(\omega)}\right)^2$ where χ denotes the susceptibility. Its comparison with experimental data from our setup in shown in Fig. 4.

In conclusion, our numerical model is shown to successfully correlate experimental SHG results with a quantitative analysis of the Si/SiO_2 interface and surface charge densities. Thus, it could pave the way for development of SHG as an efficient oxide semiconductor interface characterization tool, for this and other material systems of interest. Our efforts on another systems will also be presented.

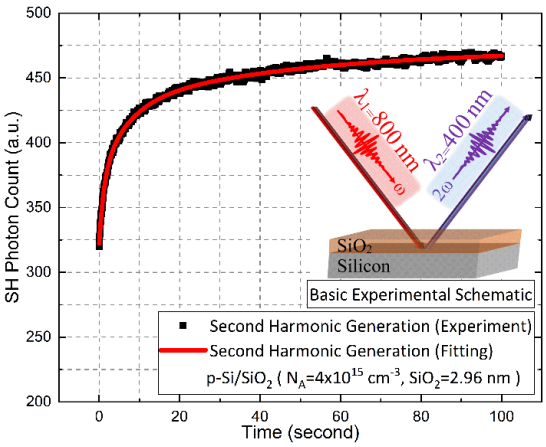


Fig.1.a: Time dependent second harmonic (SH) photon counts from experiment, fitting of semi-empirical equation to extract various trapping time constants. The schematic of SHG characterization technique captured in the inset.

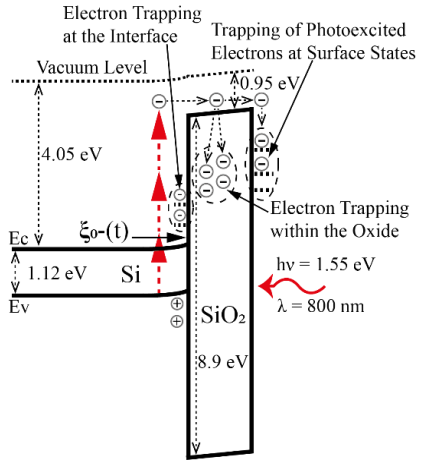


Fig.1.b: Schematic representation of the trapping of electrons excited through the multiphoton process that has been simulated here.

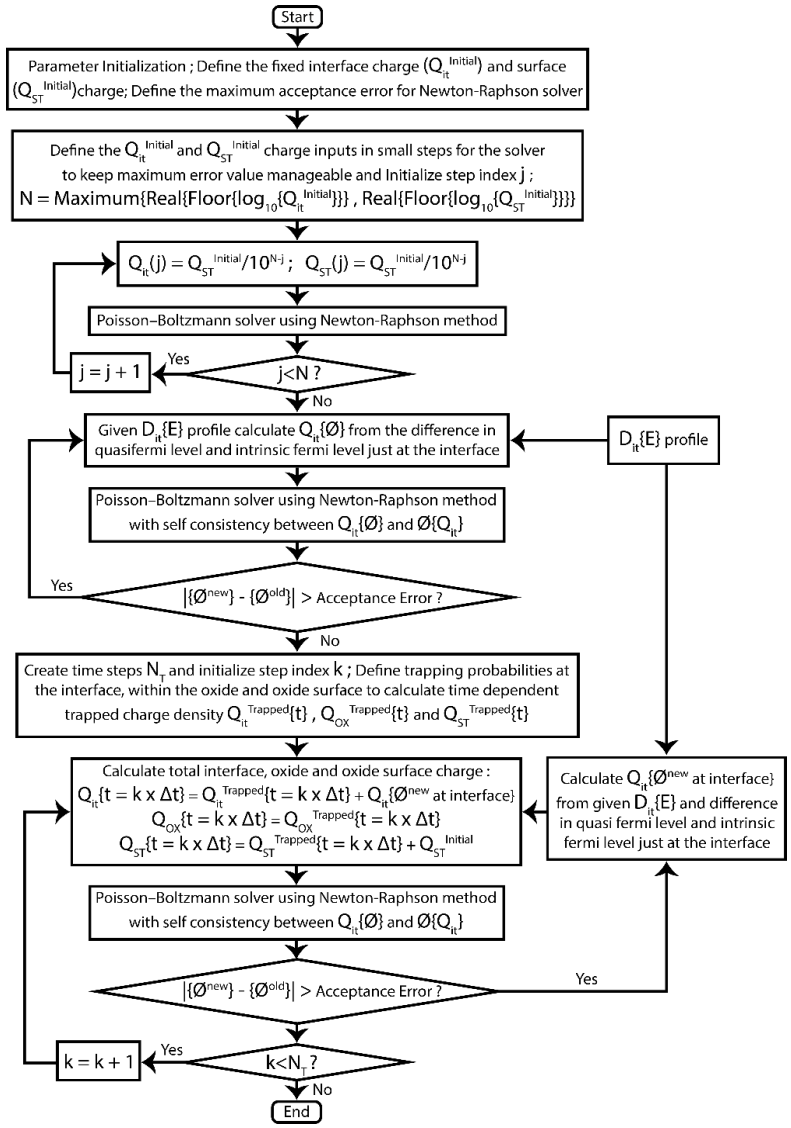


Fig.2: Flowchart of the time stepped numerical solver that captures the time dependent trapping of photo excited electrons at the Si/SiO_2 interface, within the oxide and at the surface states of SiO_2.

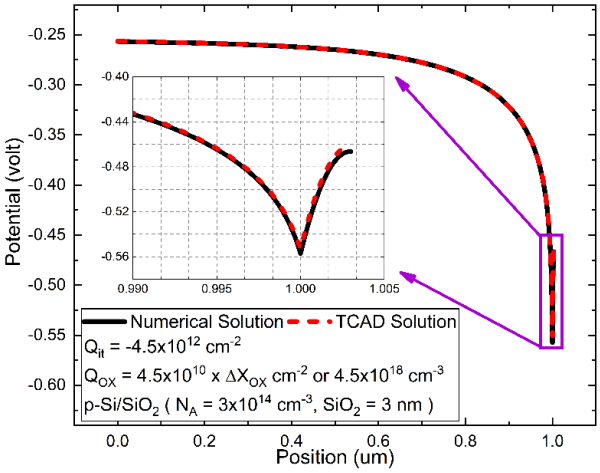


Fig.3: Comparison of simulation outcomes from Synopsys TCAD (Sentaurus Device) and our numerical solver for the validation of our solver.

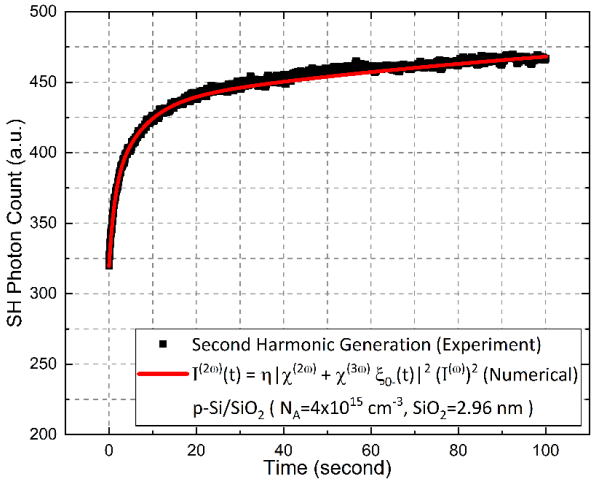


Fig.4: Comparison of experimental time dependent second harmonic photon counts and simulation result from our time stepped numerical solver.

REFERENCES:

[1] J. L. Fiore *et al.*, "Second harmonic generation probing of dopant type and density at the interface Second harmonic generation probing of dopant type and density at the Si / SiO 2 interface," vol. 041905, pp. 1–4, 2011.

[2] J. Bloch, J. Mihaychuk, and H. van Driel, "Electron Photoinjection from Silicon to Ultrathin SiO2 Films via Ambient Oxygen," *Phys. Rev. Lett.*, vol. 77, no. 5, pp. 920–923.

[3] T. Manaka, E. Lim, R. Tamura, and M. Iwamoto, "Direct imaging of carrier motion in organic transistors by optical second-harmonic generation," *Nat. Photonics*, vol. 1, no. 10, pp. 581–584, 2007.

[4] T. Katsuno, T. Manaka, T. Ishikawa, H. Ueda, T. Uesugi, and M. Iwamoto, "Current collapse imaging of Schottky gate AlGaN/GaN high electron mobility transistors by electric field-induced optical second-harmonic generation measurement," *Appl. Phys. Lett.*, vol. 104, no. 25, p. 252112, 2014.

[5] T. Katsuno, T. Manaka, T. Ishikawa, N. Soejima, T. Uesugi, and M. Iwamoto, "Three-dimensional current collapse imaging of AlGaN/GaN high electron mobility transistors by electric field-induced optical second-harmonic generation," *Appl. Phys. Lett.*, vol. 109, no. 19, p. 192102, Nov. 2016.

[6] T. Katsuno, T. Manaka, N. Soejima, and M. Iwamoto, "Direct observation of trapped charges under field-plate in p-GaN gate AlGaN/GaN high electron mobility transistors by electric field-induced optical second-harmonic generation," *Appl. Phys. Lett.*, vol. 110, no. 9, 2017.

[7] J. G. Mihaychuk, N. Shamir, H. M. van Driel, and H. M. Van Driel, "Multiphoton photoemission and electric-field-induced optical second-harmonic generation as probes of charge transfer across the Si/SiO 2 interface," *Phys. Rev. B*, vol. 59, no. 3, pp. 2164–2173, 1999.

[8] O. A. Aktsipetrov *et al.*, "dc-electric-field-induced and low-frequency electromodulation second-harmonic generation spectroscopy of Si(001)-SiO_{2} interfaces," *Phys. Rev. B*, vol. 60, no. 12, pp. 8924–8938, 1999.

[9] Y. D. Glinka *et al.*, "Characterization of charge-carrier dynamics in thin oxide layers on silicon by second harmonic generation," *Phys. Rev. B - Condens. Matter Mater. Phys.*, vol. 65, no. 19, pp. 1–4, 2002.

[10] Z. Marka *et al.*, "Band offsets measured by internal photoemission-induced second-harmonic generation," *Phys. Rev. B*, vol. 67, pp. 1–5, 2003.

[11] B. Jun *et al.*, "Charge trapping in irradiated SOI wafers measured by second harmonic generation," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 6 II, pp. 3231–3237, 2004.

[12] Binit Mallik …, "...", in preparation.

[13] D. A. Kleinman, "Theory of second harmonic generation of light," *Phys. Rev.*, vol. 128, no. 4, pp. 1761–1775, 1962.

[14] P. A. Franken and J. F. Ward, "Optical Harmonics and Nonlinear Phenomena," *Rev. Mod. Phys.*, vol. 35, no. 1, pp. 23–39, Jan. 1963.

[15] N. H. Thoan, K. Keunen, V. V. Afanas'ev, and A. Stesmans, "Interface state energy distribution and Pb defects at Si(110)/SiO2 interfaces: Comparison to (111) and (100) silicon orientations," *J. Appl. Phys.*, vol. 109, no. 1, 2011.

Modeling Thermal Effects in STT-MRAM

Tomáš Hadámek*, Wolfgang Goes‡, Siegfried Selberherr†, Viktor Sverdlov†
* Christian Doppler Laboratory for Nonvolatile Magnetoresistive Memory and Logic at the
† Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, A-1040 Wien, Austria
‡ Silvaco Europe Ltd., Cambridge, United Kingdom
Email: hadamek@iue.tuwien.ac.at

***Abstract*—In this work we employ the stochastic Landau-Lifschitz-Gilbert (sLLG) equation to explore switching in the spin-transfer torque magnetoresistive random access memory (STT-MRAM). We investigate the element size effects in the finite element method (FEM) implementation and propose an effective temperature scaling in the thermal field calculation to mitigate the dynamical behavior caused by the different element sizes.**

***Keywords*—Stochastic Landau-Lifschitz-Gilbert Equation, Spintronics, STT-MRAM, Temperature Scaling**

I. Introduction

STT-MRAM has recently gained a lot of popularity and is commercially used, however, reliable simulations tools are yet to be developed.

Due to the typical dimensions of the systems, an ab-initio simulation is not feasible and the micromagnetic approach, must be employed. The basis of the method is solving the LLG equation.

$$\frac{\partial \mathbf{m}}{\partial t} = -\gamma\mu_0 \mathbf{m} \times \mathbf{H}_{\text{eff}} + \alpha \mathbf{m} \times \frac{\partial \mathbf{m}}{\partial t} + \frac{1}{M_{\text{S}}} \mathbf{T}_{\text{S}} \qquad (1)$$

$\mathbf{m}$ stands for the normalized magnetization, M_{S} is the saturation magnetization, γ and μ_0 are the gyromagnetic ratio and vacuum permeability, $\mathbf{T}_{\text{S}}$ is a spin torque acting on the magnetization, and $\mathbf{H}_{\text{eff}}$ is the effective field. To account for the finite temperature, either a parameters scaling is need, or a random thermal field can be added to the effective field, resulting in the sLLG equation. The former method keeps the system deterministic and is not suitable to investigate the statistical behaviour of STT-MRAM switching, unlike the latter, which, however, may not correctly reproduce certain magnetic properties, like the Curie temperature. Moreover, the results depend on the element size, when a discretization method is used to solve the sLLG equation numerically [1], [2]. Here, we employ the finite element method (FEM) to solve the sLLG coupled to spin and charge transport to investigate a temperature scaling technique in order to mitigate the element size effects on the switching.

II. Methods

The random thermal field $\mathbf{H}_{\text{th}}$ added to $\mathbf{H}_{\text{eff}}$ has white noise characteristics and is uncorrelated in space and time t.

$$\langle \mathbf{H}_{\text{th},i}(t), \mathbf{H}_{\text{th},j}(t') \rangle = 2 \frac{\alpha k_{\text{B}} T_i}{\gamma \mu_0 M_{\text{S}} V_i} \delta_{ij} \delta(t - t') \qquad (2)$$

i, j are the element indices, k_{B} is the Boltzmann constant, T_i and V_i are the element temperature and volume, and δ_{ij} and $\delta(t-t')$ are the Kronecker delta and Dirac delta function, respectively. To account for the element size effects, we follow [3] and use scaled simulation temperature T_i^{sim} in (2) instead of T_i. To account for the tetrahedral element shape we sum over all surface areas $S_{i,k}^{\text{sim}}$ of the element.

$$T_i^{\text{sim}} = T_i \frac{6 S_i V_i^{\text{sim}}}{\sum_{k=1}^{4} S_{i,k}^{\text{sim}} V_i} \qquad (3)$$

The element volume and the side area are denoted by V^{sim} and S_i^{sim}. V_i and S_i are the volume and the side area of the lattice cell of the underlying microscopic model.

III. Results

We simulate a 5-layer cylindrical STT-MRAM cell with a diameter of 40 nm and employ two different meshe types. A "regular" mesh uses almost equilateral tetrahedral elements, whereas a "shallow" mesh uses elements with elongated lateral dimensions. The regular mesh has about twice as many elements in the free layer (FL), hence the volume ratio is approximately $2^{1/3}$. We run 100 switching simulations from an anti-parallel to parallel magnetization arrangement for both meshes, with and without temperature scaling using (3). The applied voltage is -1 V, the FL temperature is set to 300 K, and $a = 2.5\,\mathring{A}$, similar to the lattice constant of FeCoB.

Fig. 1 and Fig. 2 show the realizations of the switching simulations for the shallow mesh. The applied scaling accelerates the switching due to the elements being bigger than the lattice constant, effectively increasing the temperature. Fig. 3 shows a histogram of switching times obtained for the different simulations. The threshold for completing the switching is set to $0.9\ m_{\text{x}}$. Clearly, the temperature scaled simulations switch faster. This is visible in Fig. 4 where the histogram data is fitted with a Pearson's Type IV distribution [4]. Faster switching times using the scaled temperature are clearly visible for both meshes. The scaled meshes also show an improved match of the switching time distributions for the different mesh types, both the curves almost coincide, whereas the original, non-scaled simulations, show a significant difference.

IV. Conclusion

In this work, the sLLG equation is utilized to investigate switching in STT-MRAM for differed element sizes. It is shown that the results obtained with solving the bare sLLG equation show different switching times. The proposed temperature scaling method, previously used for accurate Curie temperature modeling, improves the consistency of the simulation results obtained with various mesh types.

Acknowledgment

The authors would like to acknowledge the financial support from the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association.

References

[1] V. Tsiantos *et al.*, "The effect of the cell size in Langevin micromagnetic simulations," *Journal of Magnetism and Magnetic Materials*, vol. 242-245, pp. 999–1001, 2002, proceedings of the Joint European Magnetic Symposia (JEMS'01). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304885301013658

[2] D. V. Berkov, *Magnetization Dynamics Including Thermal Fluctuations: Basic Phenomenology, Fast Remagnetization Processes and Transitions Over High-energy Barriers*. John Wiley Sons, Ltd, 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470022184.hmm204

[3] M. B. Hahn, "Temperature in micromagnetism: Cell size and scaling effects of the stochastic Landau–Lifshitz equation," *Journal of Physics Communications*, vol. 3, no. 7, p. 075009, jul 2019. [Online]. Available: https://doi.org/10.1088/2399-6528/ab31e6

[4] G. Siracusano *et al.*, "Description of statistical switching in perpendicular STT-MRAM within an analytical and numerical micromagnetic framework," *IEEE Transactions on Magnetics*, vol. 54, no. 5, pp. 1–10, 2018.

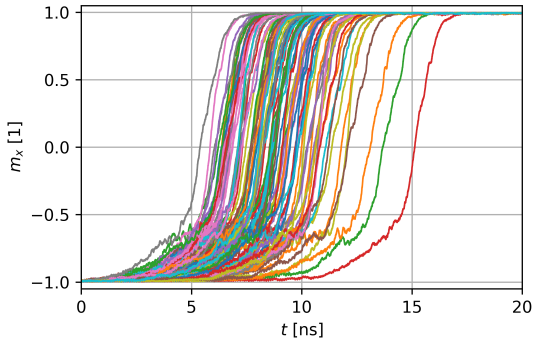


Fig. 1. Switching simulations for a shallow mesh. Thermalization of 2 ns is considered before a voltage of -1 V across the structure is applied. The structure temperature is set to 300 K.

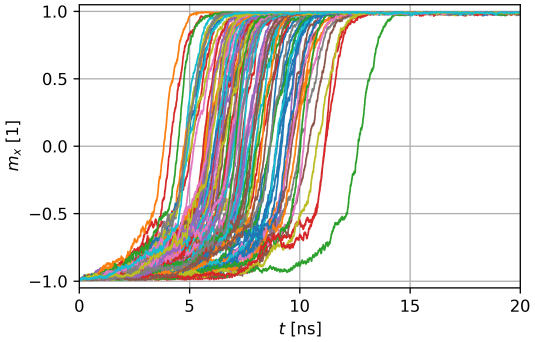


Fig. 2. Switching simulations with temperature scaling involved. Thermalization of 2 ns is considered before a voltage of -1 V across the structure is applied. The structure temperature is set to 300 K. The effectively raised temperature accounting for the elements size leads to faster switching.

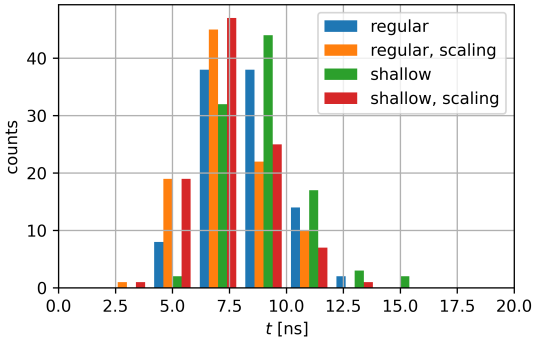


Fig. 3. Switching times for a structure with different element sizes with both, no temperature scaling and temperature scaling included. The switching threshold was set to +0.9 of the average free layer magnetization m_x

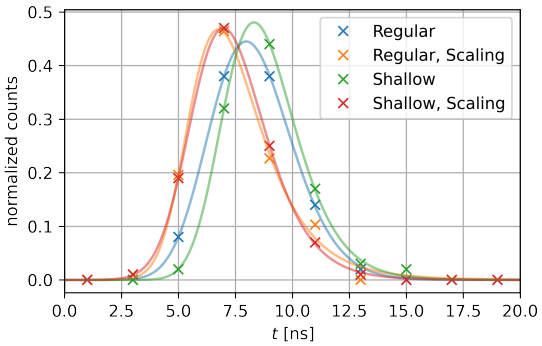


Fig. 4. Pearson Type IV distribution fitted to the switching times. The original, non-scaled simulations are slower and vary more. The scaled simulations show a very good agreement even for different element sizes.

Monolithic TCAD Simulation of Phase-Change (PCM/PRAM) + Ovonic Threshold Switch (OTS) Selector Device

M. Thesberg[1], Z. Stanojevic[1], O. Baumgartner[1], C. Kernstock[1], D. Leonelli[2], M. Barci[2], X. Wang[3], X. Zhou[3], H. Jiao[3], G. L. Donadio[4], D. Garbin[4], T. Witters[4], S. Kundu[4], H. Hody[4], R. Delhougne[4], G. S. Kar[4], M. Karner[1]

[1]Global TCAD Solutions (Vienna, Austria), [2]Huawei Technologies R&D Belguim N.V. (Leuven, Belgium), [3]HiSilicon Technologies (Shenzhen, China), [4]imec (Leuven, Belgium)

Owing to the increasing interest in the commercialization of phase-change memory (PCM) devices, a number of TCAD models have been developed for their simulation. These models formulate the melting, amorphization and crystallization of phase-change materials as well as their extreme conductivity dependence on both electric field and temperature into a set of self-consistently solved classical thermoelectric and phase-field partial-differential equations.[1] However, demonstrations of the ability of such models to match actual experimental results are rare. In addition, such PCM devices also require a so-called selector - such as an Ovonic Threshold Switching (OTS) - device[2] in series for proper memory operation[3]. However, monolithic simulation of both the PCM and OTS selector device[4] in a single simulation is largely absent from the literature, despite its value for material and design space explorations. It is the goal of this work to first characterize a PCM device in isolation against experimental data, then to demonstrate the qualitative behavior of a simulated OTS device in isolation and finally to perform a single monolithic simulation of the PCM + OTS device within the confines of the commercially available TCAD solver: GTS Framework.

PCM-Only Device: The model used for simulation is that found in Ref. 1 where the material values of that work were used except for a coefficient scaling the amorphization, nucleation and "in-mesh" and "inter-mesh" rates and the bulk conductivity of the PCM material which were treated as fittable parameters. Due to the large parameter space, 2D simulation was used. In order to determine the value of these parameters, experimental data was extracted from a mushroom cell (Figure 1) consisting of a physical vapor deposition (PVD) $Ge_2Sb_2Te_5$ (GST) layer of 50 nm thickness with a bottom "heater" TiN layer of 65 nm diameter and a top TiN electrode. A corresponding TCAD device was then created (Figures 2 and 3). 17 of the experimental devices were constructed and were then placed in series with a 750 ohm resistor and run through a "Reset"-"Set"-"Reset" (RSR) pulse sequence. This RSR sequence involves the application of an initial 2.5 V/200 ns square pulse intended to put the device into the "Reset"/amorphous state followed by an Incremental Step Pulse Programming (ISPP) scheme (Figure 3) of 1 µs square pulses of increasing voltage followed by a read pulse of 0.1 V. The resulting resistance vs. pulse voltage data is shown in Figure 4 for all 17 devices as well as the final TCAD results following a material parameter optimization search. Good agreement between TCAD and experiment is found.

OTS-Only Device: Owing to the current lack of experimental data for the OTS device at the time of publication, instead literature values for GST were used taken from Ref. 1 – which are different than those obtained from the parameterization of the PCM-only device - and modified such that the nucleation and phase growth rates were set to zero, thus ensuring that the device always re-enters the amorphous state after melting, as is optimal for an ovonic switch. To verify that this model would then exhibit the correct OTS behavior a simple triangular pulse was simulated and the current versus voltage results of which are shown in Figure 6. A clear threshold turn on at a V_{th} of ~2.5 V is observed.

PCM+OTS Device: Finally, the two materials with two different sets of parameters, separated by a thin TiN intermediate layer, were simulated as a monolithic device (Figure 7). This device was run through a triangular pulse taking it from an initial "Set" state to a "Reset" state (Figure 8). For comparison the PCM-only device was put through the same pulse. The ability of the OTS device to block flow at low voltages but permit them at high is clearly shown.

Thus a TCAD model capable of full monolithic simulation of PCM+OTS devices has been achieved and its utility as the basis for future studies has been demonstrated.

References: [1] Z. Woods et. al. IEEE Trans. on Elect. Dev. 64.11 (2017): 4466-4471, [2] G. W. Burr, et. al. J. Vac. Sci. Technol. B 32, 040802 (2014). [3] S. R. Ovshinsky, Phys. Rev. Lett. 21, 1450–1453 (1968). [4] W.-C. Chien, et. al. IEEE Trans. Elect. Dev. 65, 5172–5179 (2018)

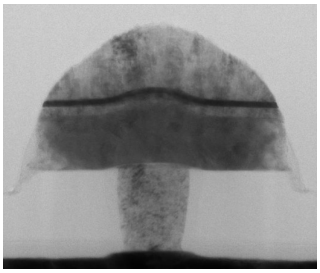


Figure 1: TEM image of PCM-only mushroom device with a 200 nm wide $Ge_2Sb_2Te_5$ (GST) PCM slab of 50 nm thickness contacted TiN cylindrical heater of 65 nmdiameter.

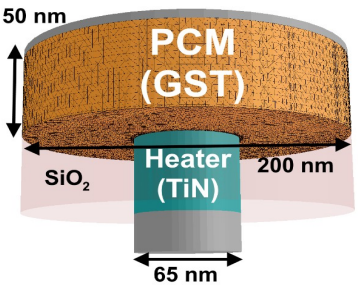


Figure 2: Device schematic of PCM-only TCAD device used for parameterization and validation of phase-change properties against experimental data.

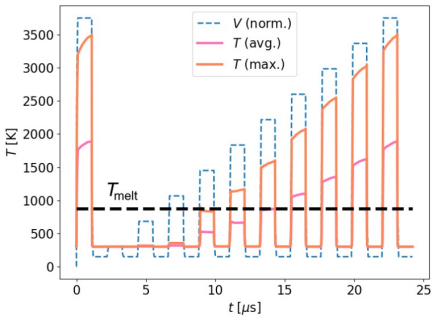


Figure 3: The normalized voltage of the Reset-Set-Reset (RSR) experimental pulse and resulting average and maximum temperature. A 750 Ohm resistor is also present in series with the device. Simulations were performed in 2D.

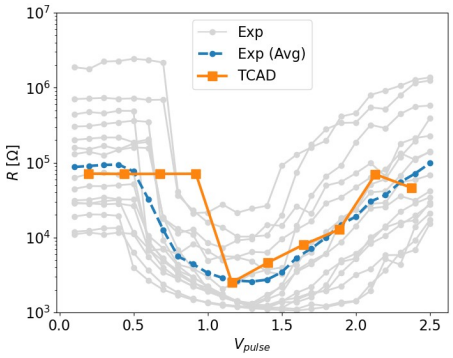


Figure 5: Experimental resistance vs. applied pulse voltage data, along with the TCAD best-fit results, for the RSR pulse (750 Ohm series resistor).

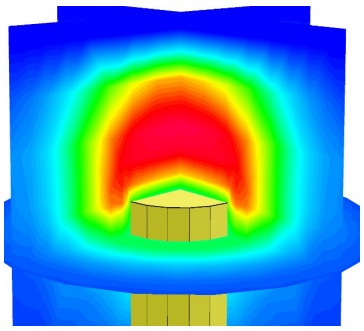


Figure 4: Illustrative image of temperature distribution during melting in TCAD simulation.

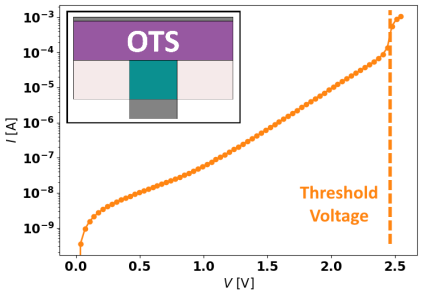


Figure 6: Current vs. voltage behavior of the OTS-only device showing a clear threshold voltage and "turn on".

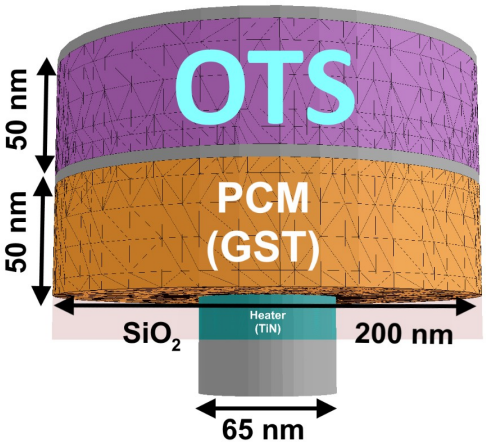


Figure 7: Device schematic of the combined PCM+OTS device.

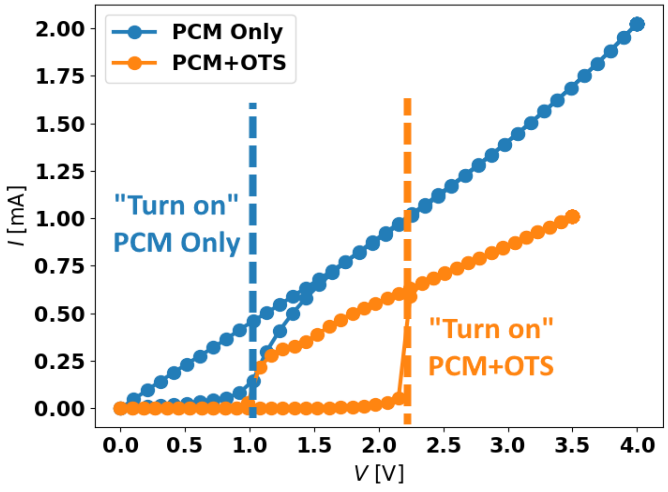


Figure 8: Current vs. voltage behavior for the PCM-only (blue) and PCM+OTS (orange) device as they are taken through a 1 μs set-reset pulse.

Performance of Vertical Gate-All-Around Nanowire p-MOS Transistors Determined by Boron Depletion during Oxidation

Chiara Rossi[1], Alexander Burenkov[1], Peter Pichler[1], Eberhard Bär[1],
Paweł Piotr Michałowski[2], Jonas Müller[3], Guilhem Larrieu[3]

[1]Fraunhofer Institute for Integrated Systems and Device Technology (IISB), Erlangen, Germany
[2] Łukasiewicz Research Network – Institute of Microelectronics and Photonics, Warsaw, Poland
[3]LAAS-CNRS, Université de Toulouse, Toulouse, France

The gate-all-around (GAA) nanowire (NW) field-effect transistor (FET) represents one of the most promising candidates to meet the increasing requirements of device performances and density. GAA-NW-FETs show excellent immunity against short-channel effects (SCE) and, thanks to their 3D layout configuration, can enhance integration and overcome some physical limitations deriving from 2D layout, such as contact placements and interconnect routing congestion [1]. Vertical GAA-NW-FETs can be fabricated using a top-down approach with conventional processes and very good control on dimensions and localization. Moreover, the gate length is given by the thickness of the deposited gate material, without the need of high-resolution lithography [2]. Silicon nanorods can be obtained by first etching suitably doped silicon substrates. Then, they are thinned using a sacrificial thermal oxidation to obtain the desired diameter, thus removing also residual damage from the etch process. During such sacrificial oxidations as well as during gate oxide growth, dopants segregate into the growing oxide and self-interstitial are injected into silicon. Unlike for the case of oxidation of a bulk planar surface, in silicon nanowires, the self-interstitials cannot diffuse efficiently into the bulk. This increases dopant diffusion and segregation loss in the nanorod while the resupply of dopants from the bulk is less effective due to device geometry. For this reason, the final doping distribution after oxidation should be carefully analyzed for a correct prediction of the electrical characteristics of the GAA-NW-FETs.

For this work, we performed 3D simulations of the full process flow and 3D device simulations with the Sentaurus Process and Device programs of Synopsys. We considered junctionless GAA-NW-FET with boron-doped vertical nanowires as described in [2]. Fig. 1 shows the simulated nanowire and the grown oxide after the sacrificial wet oxidation (850 °C, 10 minutes) and the subsequent dry gate oxidation (725 °C, 30 minutes) to obtain a gate oxide with a thickness of 4 nm [2]. Even though the oxidation rate depends on the silicon orientation and so a diamond-like shape in the nanowire cross-section could be expected, SEM images of nanowires after oxidation (Fig. 2) show a circular shape. This was also reported in literature by previous works [3-6]. The circular shape was explained by Ye *et al.* [3] considering that oxidation at edges of a surface is slower due to the increased energy barrier against oxidation of the Si-Si bond at the edges

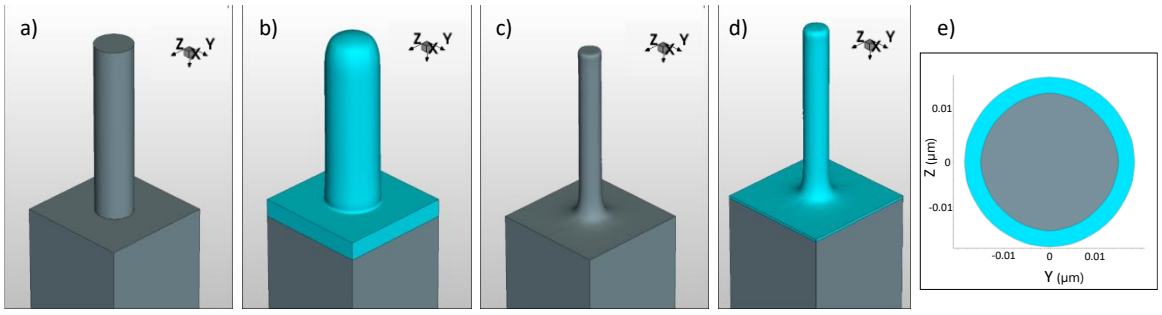


Figure 1. a) Initial structure after etching of the nanowire; b) nanowire after sacrificial thermal oxidation (oxide in light blue); c) thinned nanowire after oxide stripping; d) nanowire after dry thermal oxidation for gate oxide growth; e) cross section of the nanowire after gate oxidation: The orientation dependence of the oxidation was switched off to obtain a circular nanowire as reported in literature.

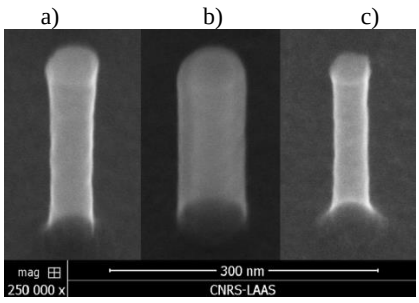


Figure 2. SEM images of nanowires fabricated at CNRS-LAAS: a) initial nanowire after etching, b) oxidized nanowire after sacrificial wet oxidation, c) nanowire after oxide stripping. The nanowire has a circular shape in the cross-section.

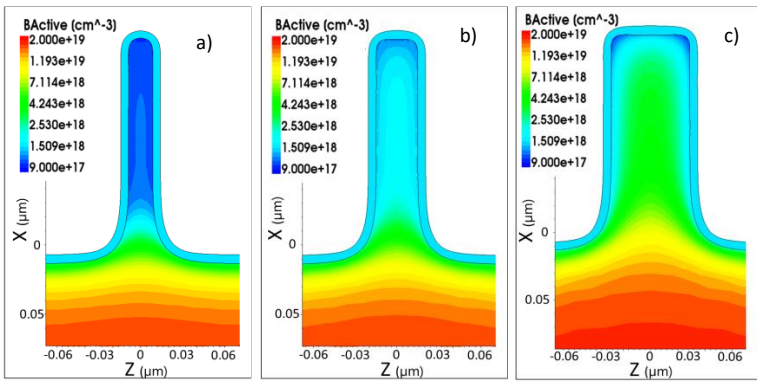


Figure 3. Electrically active doping concentration in nanowires after sacrificial wet oxidation and gate dry oxidation for three different nanowire diameter: a) 18 nm, b) 29 nm, c) 58 nm. The graphs show a 2D cut-plane of the 3D simulated structures.

so that the oxidation at the center of a surface is faster. As the oxidation proceeds, new edges are generated and the same oxidation rate is obtained all around the outer surface of nanowire, leading to a circular shape. To reproduce the experimentally observed circular shape, the orientation dependence of the oxidation rate was switched off in the simulations. Before the sacrificial oxidation, the etched nanowire has a uniform boron concentration of $2 \cdot 10^{19}$ cm^{-3}, as it was in the silicon substrate. The simulated final doping distribution after the two oxidation steps is reported in Fig. 3 for three different nanowire geometries (final diameters of 18, 29 and 58 nm). The doping concentration has been computed using the *ChargedPair* diffusion model and the three-phase segregation model [7] with the parameter set from Sentaurus TCAD *AdvancedCalibration 2021.06*. The next steps of the fabrication of GAA-NW-FETs were subsequently simulated: deposition of spacers and gate metal layer (with a thickness, thus a gate length, of 15 nm) and definition of source and drain contacts as ohmic contacts. The final simulated structure, ready for 3D device simulations, is presented in Fig. 4.

Process simulations show that the doping distribution in the nanowire changes considerably during the oxidation steps. It is lowered by about one order of magnitude in comparison with the initial dopant concentration and the decrease is more pronounced for thinner nanowires, so this effect will be even more prominent for future device generations. For smaller nanowire diameter, the surface-to-volume ratio increases and so more dopants diffuse out from silicon nanowire into the oxide. The resupply of dopants from bulk is almost suppressed in thinner nanowires due to the geometrical configuration.

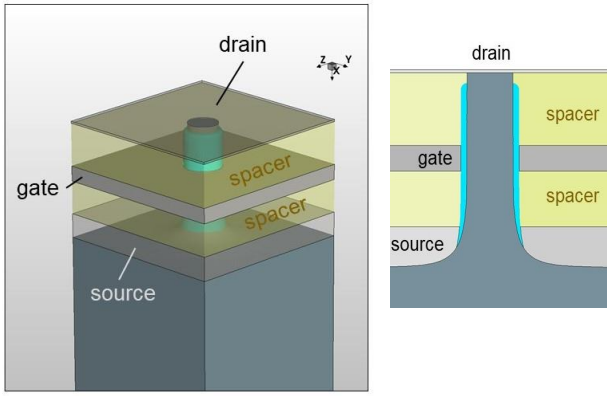


Figure 4. Final 3D structure of GAA-NW-FET at the end of process simulation, on the right: cross section

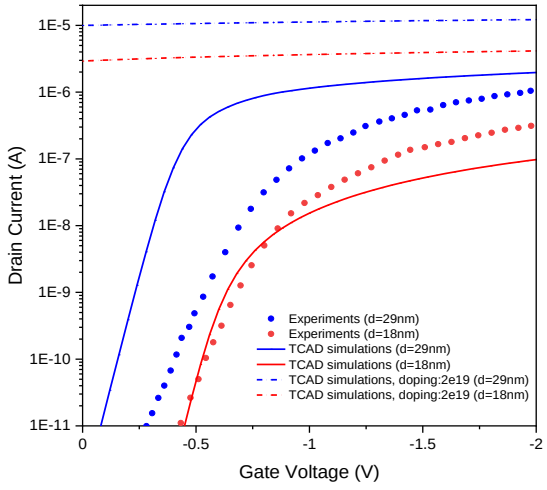


Figure 5. Transfer characteristics of GAA-NW-FET (V_d=-0.1V). Simulated results are with lines, measured data [2] with symbols. Dashed lines refer to simulations without oxidation simulation (doping in NW: $2 \cdot 10^{19}$ cm^{-3})

In junctionless transistors, the evaluation of depletion region is of high importance for a correct prediction of the electrical characteristics. The results of the electrical device simulations are shown in Fig.4 in comparison to the measured data from [2]. Without taking into consideration the boron loss in the nanowires during oxidation, the switch-off of the junctionless GAA-NW-FET cannot be reproduced in simulations. When simulating the nanowire with a diameter as it was found in the experiment, but with a boron concentration of $2 \cdot 10^{19}$ cm^{-3}, as it was initially in the nanorod, a full depletion of the channel could not be achieved for any gate voltage considered. In the result, the source-to-drain current is then almost independent on gate voltage, as shown in Fig. 5 by the dot-dashed lines.

With the full process simulation, considering the boron depletion due to segregation towards silicon dioxide during the oxidation, we could obtain typical pMOSFET characteristics (solid lines in Fig. 5) with threshold voltages similar to those obtained in experiment, even though a perfect fit was not achieved yet. For a better agreement, an improved description of the doping profiles in the nanowires is required. With boron segregation having an important impact, we also used different segregation model parameters set, for example, the one reported in [8] extracted for a similar temperature range. Unfortunately, this did not significantly improve the simultaneous reproduction of the transfer characteristics for the nanowire diameters available (not shown). Moreover, to be able to discern between the effects of the dopant distribution and device-related parameters like interface charges, complementary information about the dopant distributions in the nanorods is needed. To this end, an advanced methodology of SIMS characterization has been developed and first results will be available to be shown at the conference.

In conclusion, this work presents full 3D TCAD process and device simulations of vertical GAA-NW-FETs. The doping distribution in the nanowire is deeply affected by thermal oxidation, thus a full 3D simulation of the oxidation steps is imperative. Device simulations could qualitatively reproduce the experimentally observed switch-off characteristics of GAA transistors only if the effect of the strong boron depletion in nanowires during the oxidation was accounted for. However, TCAD model parameters for diffusion and segregation still need to be improved for a better agreement to the measured electrical data.

References

[1] A. Veloso *et al.*, "Vertical Nanowire FET Integration and Device Aspects", ECS Trans. 72(4), 31 (2016).

[2] Y. Guerfi and G. Larrieu, "Vertical Silicon Nanowire Field Effect Transistors with Nanoscale Gate-All-Around", Nanoscale Research Letters 11, 210 (2016).

[3] S. Ye *et al.*, "Precise Fabrication of Uniform sub-10-nm-Diameter Cylindrical Silicon Nanopillars via Oxidation Control", Scr. Mater. 198, 113818 (2021).

[4] P.-F. Fazzini *et al.*, "Modeling Stress Retarded Self-Limiting Oxidation of Suspended Silicon Nanowires for the Development of Silicon Nanowire-Based Nanodevices", J. Appl. Phys. 110, 033524 (2011).

[5] M. Uematsu *et al.*, "Two-Dimensional Simulation of Pattern-Dependent Oxidation of Silicon Nanostructures on Silicon-on-Insulator Substrates", Solid-State Electron. 48, 1073 (2004).

[6] X.-L. Han *et al.*, "Modelling and engineering of stress based controlled oxidation effects for silicon nanostructure patterning", Nanotechnology 24(49), 495301 (2013)

[7] F. Lau *et al.*, "A Model for Phosphorus Segregation at the Silicon-Silicon Dioxide Interface", Appl. Phys. A 49, 671 (1989).

[8] S. Koffel *et al.*, "On an Improved Boron Segregation Calibration from a Particularly Sensitive Power MOS Process", Phys. Status Solidi C 11(1), 12 (2014)

Polarization Switching Characteristics in AFE/FE Double–Layer Devices

Mengqi Fan, Fei Liu, Xiaoyan Liu
School of Integrated Circuits, Peking University, Beijing, 100871, China. Email: xyliu@pku.edu.cn

Introduction

FE/AFE/FE stack-based multi-bit memory has recently been proposed, which shows improved device-to-device variation control [1]. To ensure non-volatile storage, switching current of the same polarity should appear at the same voltage side and to reduce states overlap, large separation between these current peaks is favorable [2]. The interplay between FE and AFE layers hasn't been studied which influences the switching current profile. We investigate the polarization switching characteristics in AFE/FE double layers and analyze their differences from stand-alone devices due to the interplay between the layers. The impacts of spontaneous polarization charges and film thicknesses on the current shift are evaluated, which provides design guidance for these multi-layer devices.

Method

Fig.1 shows the studied double-layer device. Each layer is composed of 500 independently switching grains. The Monte-Carlo based nucleation-limited switching (NLS) [3][4] is used to model the polarization switching in FE and AFE layers, in which the state of FE grain is either +1(↑) or −1(↓), while the AFE grain takes on one of the three values +1, 0, −1. The electrostatics of the dielectric stacks are solved consistently with the NLS models. Charge trapping is not considered in this work. The adopted AFE/FE parameters are listed in Table.1.

Discussion

We investigate the switching characteristics in AFE(5nm)/FE(2nm) double-layer device with different spontaneous charges. The applied voltage sweeps from 4.5V to −4.5V and then back. Fig.2 shows the typical polarization-electric field (PE) loops and current response in the three cases, where Ps_{FE} is equal to, smaller or larger than Ps_{AFE}. The responses of AFE stand-alone devices are also plotted as reference. In all three cases, the 2nd current peaks shift leftwards, although to a different extent. In Case I, the 1st current peaks on the upward branch barely budge compared to stand-alone AFE devices, while in Case II and Case III, they shift left and right respectively. The responses on the downward branches are symmetrical so in the following discussion, we will discuss the upward branches.

First, we start with a detailed example of Case I with $Ps_{AFE} = Ps_{FE} = 10\ \mu C/cm^2$. Fig.3(a) shows the normalized polarization in FE and AFE layer respectively, along with those of the stand-alone devices. Four stages during the upward switching are highlighted. At Stage①, all FE and AFE grains are polarized down and there's no depolarization field. The grains in double layers see the same environment as in the stand-alone devices. Since the AFE grains generally have smaller switching time constant τ, AFE switching (from ↓ to 0) occurs earlier than the FE switching. Once the AFE grains switch to zero-state, the polarization charges in the underlying FE grains are not balanced, giving rise to large depolarization field E_{dep}. Therefore, in Stage② the depolarization of FE and AFE go hand in hand. Compared with stand-alone FE devices, the E_{dep} accelerates the FE depolarization in the double-layer device. Note that, the depolarization of AFE grain is stable due to the large backswitching barrier, whereas the FE grains oscillate between the two polarized states. When the external bias field E_{ext} changes polarity, the increasing E_{ext} helps the FE grains to stay polarized-up in Stage③. In Stage④, the positive polarization charges assist in faster AFE switching.

With larger FE spontaneous polarization charges $Ps_{FE} = 20\ \mu C/cm^2$, there are some differences in Stage③④ as illustrated in Fig.3(b). The AFE switching (from 0 to ↑) does not wait until the FE switching stabilized. Instead, the FE polarization charges are sufficiently large to prompt the AFE grains to switch up even when the E_{ext} is weak, which in return stabilizes the FE grains. This is evidenced in Fig.4 which shows the grain pattern compositions of the double-layer device. For the case of Ps = 10 μC/cm^2, there's a moment (around 30 μs) when the 0/↑ grains account for more than 95% of the AFE/FE grains. In contrast, for the case of Ps = 20 μC/cm^2, the premature ending of the increase of 0/↑ grains agrees with the early onset of ↑/↑ grains. The FE and AFE polarization switching are closely entwined in Fig.3(b) and hence the switching current peaks are sharper than those in Fig.3(a). On the other hand, if Ps_{FE} is very small, its E_{dep} is not large enough and the FE depolarization current might be separated from the AFE depolarization current (i.e. 1st current peak) as shown by the hump in Fig.2(e).

As for Case II, the main difference from Case I lies in Stage①, where the under-compensation of the AFE polarization charges, leads to the expedited AFE switching compared with stand-alone AFE devices. Conversely, in Case III, the AFE switching is delayed.

In all three cases, the shift of the current peaks can be estimated using the following equation [5]:

$$\mathrm{V_{FE}} + \mathrm{V_{AFE}} = \frac{\mathrm{Q} - \mathrm{P_1}}{C_{\mathrm{FE}}} + \frac{\mathrm{Q} - \mathrm{P_2}}{C_{\mathrm{AFE}}} = 0$$

where Q is the charges on electrodes; P_1 and P_2 are the polarization in FE and AFE respectively. C_{FE} and C_{AFE} are their capacitance. Then, we have:

$$\Delta E_C = \mathrm{V_{AFE}}/T_{AFE} = (P_1 - P_2)/(\varepsilon_{FE}T_{AFE}/T_{FE} + \varepsilon_{AFE}) \quad (1)$$

As shown in Fig.5 and Fig.6, the calculated results agree well with the simulated results. Both $\Delta E_{C1}(\propto Ps_{FE} - Ps_{AFE})$ and $\Delta E_{C2}(\propto Ps_{FE})$ increase with decreasing thickness ratio T_{AFE}/T_{FE}.

Conclusion

We investigate the switching characteristics of AFE/FE double layer devices using NLS model. It is found that AFE and FE switching are more closely coupled in devices with larger Ps_{FE}. The switching current shifts and their dependences on spontaneous polarization and film thicknesses are given, which may serve as guidance for device design.

Reference

[1] Y. Xu, *et al*., IEDM, 2021. pp.126-129. [2] K. Ni et al., IEDM, 2019. pp. 28.8. 1-28.8.4 [3] Y.-C. Chen *et al*., IEDM, 2021. pp.338-341. [4] C. Alessandri, *et al*., IEDM, 2018. pp.368-371. [5] T. P. Ma *et al*., EDL, 2002. vol. 23, no. 7, pp. 386-388.

For each grain, the switching probability:

$$P_{sw}(t) = 1 - \exp\left[-\left(\frac{t}{\tau}\right)^n\right]$$

where: $\tau(E) = \tau_0 \exp[(E_a/|E - E_b|)^\alpha]$ for AFE grain

$\tau(E) = \tau_0 \exp[(E_a/|E|)^\alpha]$ for FE grain

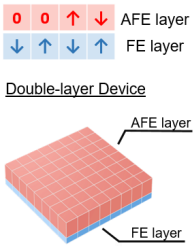


Table 1. NLS Parameters

	AFE
E_a	2.3 ± 0.32 MV/cm
E_b	2.1 ± 0.41 MV/cm
τ_0	73 ns
	FE
E_a	1.83 ± 0.43 MV/cm
τ_0	1203 ns
	Common
α	4.11
n	1.02

Fig.1 Structure of the AFE/FE double layer device. The adopted NLS parameters are summarized in Table.1.

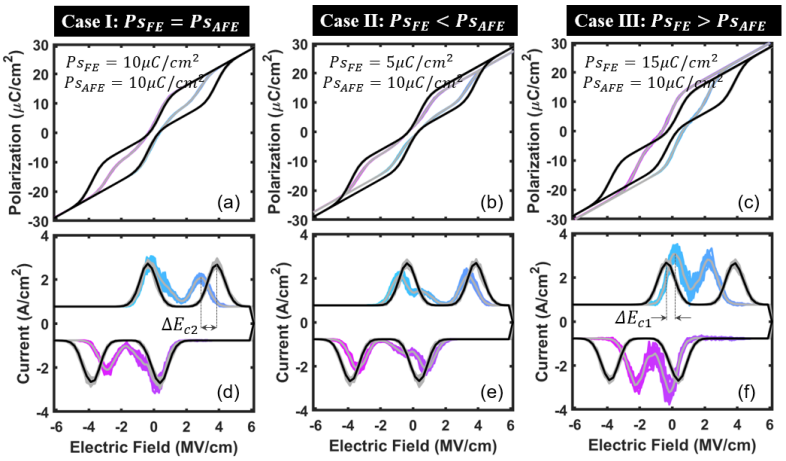


Fig.2 Simulated PV loops (a-c) and switching current (d-f) of the AFE/FE double layers (colored) and the reference stand-alone AFE characteristics (black). $f = 10$ kHz. Traces of 20 devices are shown in each case.

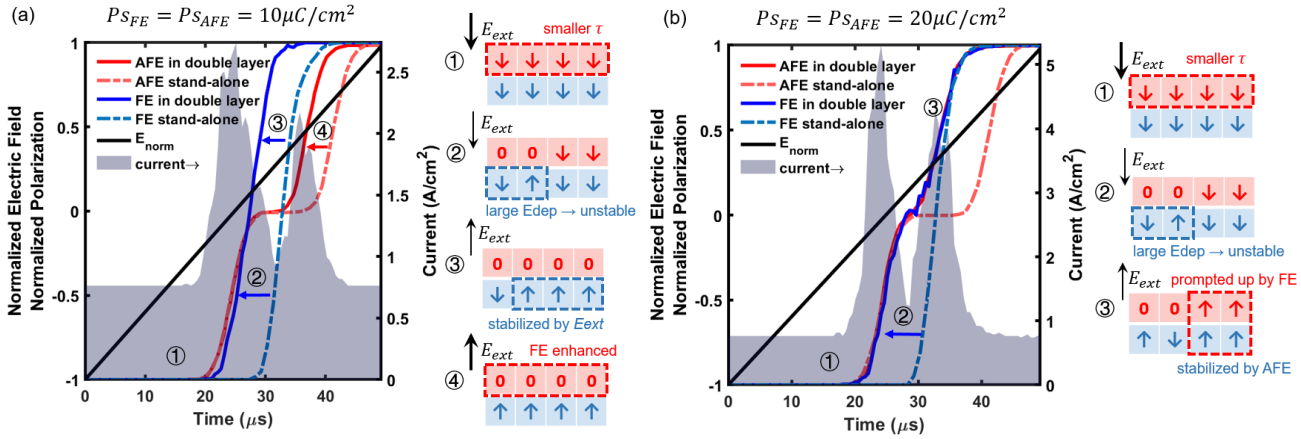


Fig.3 Evolution of normalized polarization in FE and AFE layer in the Case I double-layer device (solid), compared with those in stand-alone devices (dashed). The outline of the shaded area represents the switching current. (a) With small Ps_{AFE}, AFE switching (stage④) lags behind FE switching (stage③); (b) With large Ps_{AFE}, FE and AFE polarization switching are closely entwined in Stage ③, resulting in sharper switching current peaks.

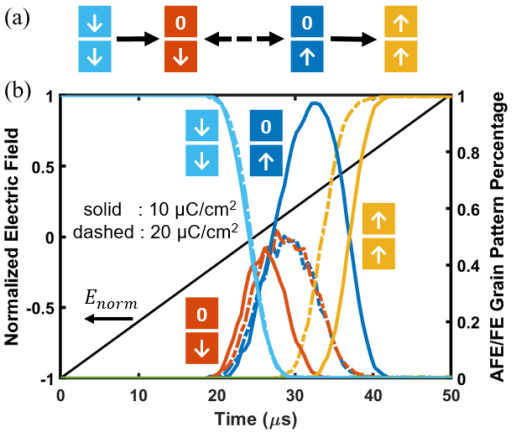


Fig.4 (a) The transition path of the polarization pattern for AFE/FE grain. (b) The composition evolution of AFE/FE grain patterns for devices with $Ps_{FE} = Ps_{AFE} = 10\mu C/cm^2$ (solid) and $Ps_{FE} = Ps_{AFE} = 20\mu C/cm^2$ (dashed).

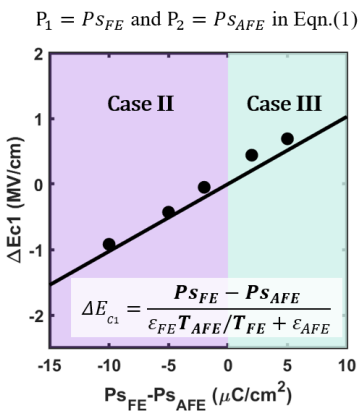


Fig.5 The linear dependence of 1st current peak shift ΔE_{C1} on $Ps_{FE} - Ps_{AFE}$.

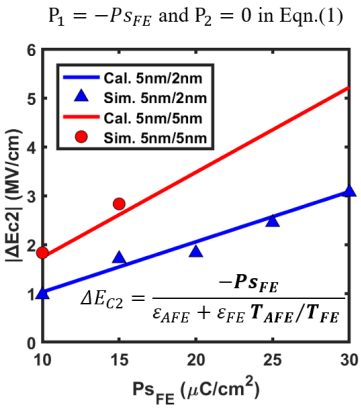


Fig.6 The linear dependence of 2nd current peak shift ΔE_{C2} on Ps_{FE}. Increase thickness ratio T_{AFE}/T_{FE} results in increase ΔE_{C2}.

Scattering matrix-based low computational cost model for the device and circuit co-simulation of phosphorene tunnel field-effect transistors

Kosuke Yamaguchi and Satofumi Souma†
Department of Electrical and Electronic Engineering, Kobe University, Kobe 657-8501, Japan
†email: ssouma@harbor.kobe-u.ac.jp

***Abstract*—We present a comprehensive device and circuit co-simulation scheme for phosphorene tunnel field effect transistors, where we employ the multiband effective mass equations with the effective masses extracted from the atomistic tight-binding model, and then calculate the current flowing through the device based on the Landauer-Büttiker formula. Here the band-to-band transmission probabilities are calculated efficiently using the scattering matrix combination method assuming the simplified potential profile, enabling the significantly low computational cost with keeping the accuracy. The obtained drain current $I_{\rm D}(V_{\rm G}, V_{\rm D})$ as a functions of the gate voltage $V_{\rm G}$ and the drain voltage $V_{\rm D}$ is tabulated for various device parameters, and are used in the SPICE circuit simulations via the Verilog-A interface quickly, demonstrating that our proposed scattering matrix-based model is beneficial for the device and circuit co-simulations.**

I. Introduction

In recent years, with the expansion of the applications in various fields of LSI technology such as neuromorphic computing and quantum computing, the demand for miniaturization and higher performance of field-effect transistors (FETs) has been further increasing. Although the miniaturization of FETs to nanometer scale causes serious increase in off-leakage current, it is expected that the superiority of the electrostatic control characteristics in two-dimensional semiconductors is one of the key strategies in the future electronics. Among two-dimensional semiconductors, phosphorene, which is one of the allotropes of phosphorus, is attracting attention as a material having an appropriate band gap and high mobility [1].

From the viewpoint of new device structures/operating principles, on the other hand, tunnel field effect transistor (TFET) has been expected as a key technology [2], where the use of band-to-band quantum tunnel effect for the switching mechanism is expected to improve the switching performance (steeper sub-threshold swing) compared to the conventional FET. Therefore it is expected that the TFETs using phosphorene as a channel material can be one of the promising technology in the future. However, while there are many studies on the electrical conduction characteristics of phosphorene TFETs, the expected superiority when used as a component of the circuit has not been fully investigated so far. Therefore, in this study we comprehensively discuss the phosphorene TFETs from the device simulation and circuit simulation viewpoints.

978-1-6654-4200-8/21/$31.00 ©2021 IEEE

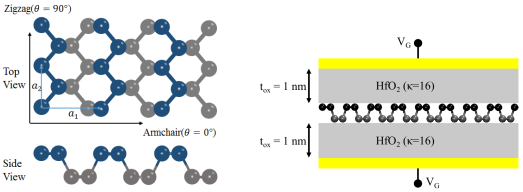


Fig. 1. Schematic illustrations of phosphorene crystal structure and phosphorene-based FET.

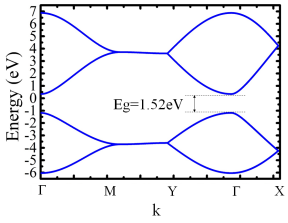


Fig. 2. Electronic band structure of phosphorene, where the bandgap is estimated as 1.52 eV.

II. Model and method

Electronic properties of phosphorene shown in Fig. 1 (left) can be analyzed effectively using the tight-binding (TB) model [3], where the band structure is calculated by solving the eigenvalue problem $H(\boldsymbol{k})\,|\psi_{l\boldsymbol{k}}\rangle = E_l(\boldsymbol{k})\,|\psi_{l\boldsymbol{k}}\rangle$ with the $\boldsymbol{k} = (k_x, k_y)$ dependent 4×4 TB Hamiltonian $H(\boldsymbol{k})$. Then the electronic band structure is obtained as shown in Fig. 2, where the band gap energy is estimated to be 1.52 eV. Then electronic transport properties of phosphorene TFET can be modeled by the following two-band effective mass equation model [4].

$$\left[-\frac{\hbar^2}{2m^*_{c,x}}\frac{d^2}{dx^2} + E_c + V(x) + \varepsilon_c\left(k_y\right)\right]\psi_c(x) + C\frac{dV(x)}{dx}\psi_v(x) = E\psi_c(x) \quad (1)$$

$$\left[-\frac{\hbar^2}{2m^*_{v,x}}\frac{d^2}{dx^2} + E_v + V(x) + \varepsilon_v\left(k_y\right)\right]\psi_v(x) + C\frac{dV(x)}{dx}\psi_c(x) = E\psi_v(x), \quad (2)$$

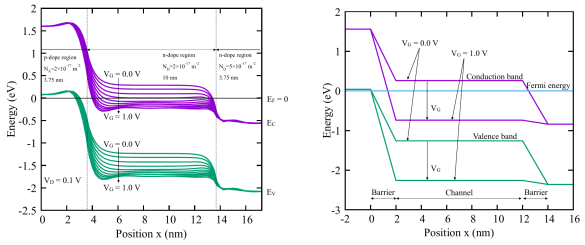


Fig. 3. (Left) Band edge profiles obtained by using NEGF method. (Right) Simplified band edge profiles assumed in this study.

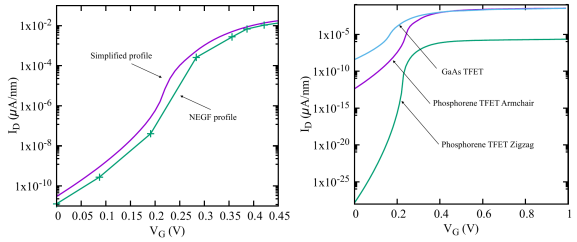


Fig. 4. (Left) Comparison of $I_{\rm D}$-$V_{\rm G}$ characteristics under the NEGF and simplified band edge profiles. (Right) Comparison of GaAs TFET, phosphorene TFET with the armchair and zigzag transport directions.

where $\varepsilon_c(k_y) \equiv \hbar^2 k_y^2/(2m^*_{c,y})$ and $\varepsilon_v(k_y) \equiv \hbar^2 k_y^2/(2m^*_{v,y})$ are the transverse energy dispersions for the conduction and valence bands, respectively, and the effective masses for the conduction (valence) band are $m^*_{c(v),x}$ and $m^*_{c(v),y}$ for x and y directions. We choose the x direction to be the transport direction, which is either the armchair or zigzag direction. The conduction band effective masses are extracted from the TB band structure as $m^*_c = 0.152m_0$ and $0.763m_0$ for armchair and zigzag directions, respectively, and corresponding valence band values are $m^*_v = 0.203m_0$ and $1.526m_0$, respectively. The parameter C is the inter-band coupling constant [4].

III. Results and Discussions

Based on the above mentioned two-band effective mass equation model, we calculate the current flowing through the device based on the Landauer-Büttiker formula, where the band-to-band transmission probabilities are calculated using the scattering matrix combination method assuming the simplified potential profile. In Fig. 3 (left) we show the band edge profiles calculated precisely using the non-equilibrium Green's function (NEGF) method. In our purpose, we approximate the band edge profiles in Fig. 3 (left) by the simplified one plotted in Fig. 3 (right) to reduce the computational time. In spite of the drastically simplified band edge profiles, we have succeeded to obtain the $I_{\rm D}$-$V_{\rm G}$ curves semi-quantitatively close to that obtained by NEGF method as shown in Fig. 4(left). We next have made the comparisons of $I_{\rm D}$-$V_{\rm G}$ characteristics for GaAs TFET, phosphorene TFET with the armchair and zigzag transport directions, as shown in Fig. 4 (right). The channel length is fixed at $L_{\rm ch} = 10$ nm. Here we can see that the on-current of phosphorene TFET with the armchair transport

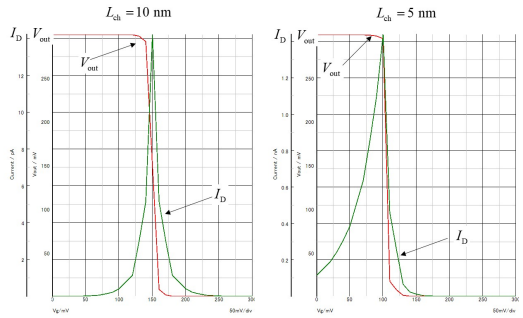


Fig. 5. Output voltage and through current of CMOS logic circuit for the channel length $L_{\rm ch} = 10$ nm (left) and 5 nm (right). Even for the extremely short channel case the expected inverter operation was obtained, although the through current was significantly increased.

directions is comparable to that of GaAs TFET, while the off-leakage current is several magnitude lower in the former than in the latter, suggesting the superiority of phosphorene TFET.

The obtained values of the drain current $I_{\rm D}(V_{\rm G}, V_{\rm D})$ as a functions of the gate voltage $V_{\rm G}$ and the drain voltage $V_{\rm D}$ have been tabulated for various parameters, and then used in the SPICE-based circuit simulations via the Verilog-A interface. For this circuit simulation purpose we employed the simulation software SIMetrix. In Fig. 5 we show the simulation results of the output voltage and through current of CMOS inverter composed of phosphorene TFET with the armchair transport directions for the channel length $L_{\rm ch} = 10$ nm (left) and 5 nm (right). As shown in Fig. 5, even for the extremely short channel case the expected inverter operation was obtained, although the through current was significantly increased.

IV. Conclusion

We presented a comprehensive simulation study of phosphorene TFETs from the device and circuit simulation viewpoints. We employed the multiband effective mass equations with the effective masses extracted from the atomistic TB model, and then calculated the current flowing through the device based on the Landauer-Büttiker formula, where the band-to-band transmission probabilities have been calculated using the scattering matrix combination method assuming the simplified potential profile to reduce the computational time. The obtained values of the drain current $I_{\rm D}(V_{\rm G}, V_{\rm D})$ as a functions of the gate voltage $V_{\rm G}$ and the drain voltage $V_{\rm D}$ have been tabulated for various parameters, and the SPICE-based circuit simulations via the Verilog-A interface has been successfully performed, demonstrating that even for the extremely short channel case the expected inverter operation can be obtained, although the through current is significantly increased.

References

[1] H. Liu, A. T. Neal, Z. Zhu, Z. Luo, X. Xu, D. Tománek, P. D. Ye, ACS Nano **8**, 4033 (2014).
[2] F. Liu, Q. Shi, J. Wang, and H. Guo, J. Appl. Phys. **107**, 203501 (2015).
[3] A. N. Rudenko and M. I. Katsnelson, Phys. Rev. B **89**, 201408(R) (2014).
[4] O. Morandi, and M. Modugno, Phys. Rev. B **89**, 235331 (2005).

Sensitivity enhancement in OCD metrology by optimizing azimuth angle based on the RCWA simulation

Hyunsuk Choi[1], Kwangseok Lee[1], Jiseong Doh[1], Jaehoon Jeong[1],
Minseok Kim[2], Yeonjeong Kim[2], Jongchul Kim[2], Hyung Keun Yoo[2], Dae Sin Kim[1]

[1]Computationl Science and Engineering Team, [2]Foundry Metrology & Inspection Technology Team
Samsung Electronics Co., Ltd., Gyeonggi-do 18448, Republic of Korea.
* Email Address: hs9207.choi@samsung.com (H. Choi).

Abstract: Monitoring process using optical critical dimension (OCD) metrology has been performed with a pre-fixed azimuth angle for all modules. In this study, we suggest optimal azimuth angle for 2 front-end modules of FinFET device based on the RCWA simulation. The sensitivity at the optimal point was improved around 30% for two modules compared to that of pre-fixed angle. Finally, the result will be compared with experimental data obtained by OCD measurement.

1. Introduction

For decades, Moore's law, the density of integrated circuit (IC) doubles every two years, has been well applied [1]. However, it seems that the gate length of IC finally reached its limit of a few nanometer-scale [2]. With down-sized gate length, the modern semiconductor structure becomes more complicated and 3-dimensional. Thus, monitoring critical dimensions (CDs) of nanostructure in the fabrication process becomes more important [3-4]. OCD metrology, based on the scatterometry, is one of the powerful technique to investigate such complex structures [5-6]. The multi-parameter, fast and non-destructive character of OCD make it a proper tool to investigate stacked nanostructures, such as FinFET or Gate-All-Around (GAA) device, which are invisible to top-down imaging methods.

Monitoring fabrication process using OCD has been performed for all module with a pre-fixed incident and azimuth angle, typically 65 and 45 degree, respectively. It is hard to modify incident angle because resetting the measuring equipment can affect established condition. On the other hand, the azimuth angle can be easily tuned by changing the wafer rotation. In this study, we suggest optimized azimuth angle for 2 monitoring steps to achieve higher sensitivity based on the rigorous coupled wave analysis (RCWA) simulation. The simulation study is performed for FinFET modules and will be applied for GAA device modules. Finally, the result will be compared with experimental data obtained by OCD measurement.

2. Experiment

A. Sensitivity In this study, sensitivity is defined as the difference in Mueller components when a structure parameter is changed from its possible minimum to maximum value, as shown in Eq. (1). Here, $\Psi_{\min}$ and $\Psi_{\max}$ represents Mueller spectrum of minimum and maximum CD, respectively, and i denotes the upper-triangle terms of Mueller Matrix, except for mm11. N is the number of Mueller components and λ is the normalization factor, 2 for Mueller elements. The components are averaged for the wavelength, x.

$$S = \frac{1}{N}\sum_i^N \frac{1}{\lambda|x|}\int dx\left|\Psi_{max}^i(\vec{x}) - \Psi_{min}^i(\vec{x})\right| \quad \text{Eq. (1)}$$

B. Experimental details Figure (1) shows the schematic view of RCWA simulation for the FinFET structure. Here, azimuth angle is defined as the angle between y-axis (normal vector of FinFET front surface) and the incident plane. In the simulation, azimuth angle is changed from 0 to 90 degree with the step size of 5 degree. For each step, a structure parameter is set to be its minimum and maximum value and their Mueller spectra are obtained by RCWA simulation, respectively. The incident angle is fixed as 65 degree. Then, the sensitivity is calculated for each step, and the optimal azimuth angle is determined as the point where sensitivity shows its maximum. The simulation was performed for 2 front-end modules of FinFET device. For each module, Fin width and Gate width was set to be the critical structure parameter, respectively.

3. Results and analysis

Figure (2) shows typical Mueller spectra obtained by RCWA simulation at the front-end module of FinFET fabrication process. In the simulation, Fin width is changed in a few nanometer range and other structure parameters are fixed. For each Mueller component, it shows spectrum of azimuth angle from 0 (purple) to 90 (red) degree. Based on Eq. (1), sensitivity is calculated for each azimuth angle and the result is shown in Figure (3). In the figure, sensitivity increases up to 0.095 with increasing azimuth angle. For the azimuth angle larger than 80 degree, sensitivity decreases as the structural symmetry recovers from the viewpoint of light. At the pre-fixed 45 degree, the value of sensitivity is 0.071, which is 0.024 smaller than the maximum value. Typically, around 0.01 difference in sensitivity is observable change in OCD measurement. Thus, 80 degree will be the optimal azimuth angle for distinguishing Fin width. Same analysis applied for the other front-end module of FinFET, where the gate width was changed. The sensitivity for gate width has its maximum value at 20 degree, as shown in Figure (4). This result is consistent with our common sense that Fin and gate width can be easily seen at the side view and front view of FinFET structure, respectively. The detailed results are shown in Table 1. Comparing to the pre-fixed azimuth

angle, sensitivity at the optimal point is enhanced 33.8% and 28.6% for two front-end modules, respectively.

4. Conclusion

In this study, we suggested optimal azimuth angle for 2 front-end modules of FinFET device based on the RCWA simulation. At the optimal angle, sensitivity was enhanced around 30% compared to the pre-fixed angle. For the future step, we will perform the simulation for GAA modules and the result will be compared with OCD measurement data.

5. Reference

[1] Mack, C. A. *IEEE Trans. Semicond. Manuf.* **24**, 202–207 (2011)
[2] Khan, H. N., Hounshell, D. A. & Fuchs, E. R. H. *Nat. Electron.* **1**, 14–21 (2018)
[3] Orji, N.G., Badaroglu, M., Barnes, B.M. *et al. Nat Electron* **1,** 532–547 (2018)
[4] Markov, I. L. *Nature* **512**, 147–154 (2014)
[5] E. Garcia-Caurel, et al. Applied Spectroscopy 67, 1-21 (2013)
[6] Ray J. Hoobler, Ebru Apak, Proc. SPIE 5256, 23rd Annual BACUS Symposium on Photomask Technology, (2003)

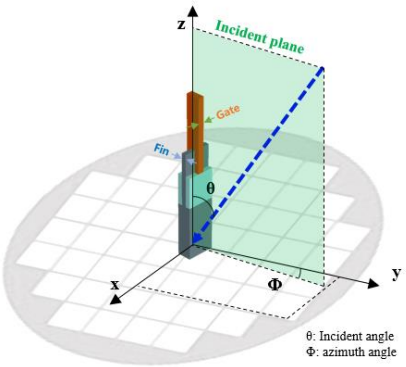


Fig. 1. Schematic view of RCWA simulation for FinFET device

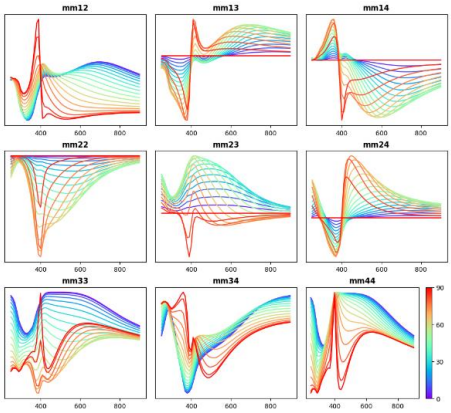


Fig. 2. Typical Mueller Matrix elements of FinFET device obtained by RCWA simulation.

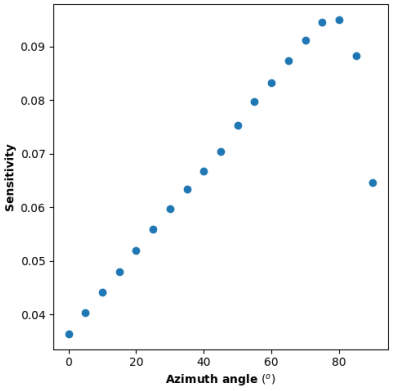


Fig. 3. Sensitivity plot for front-end module1 of FinFET.

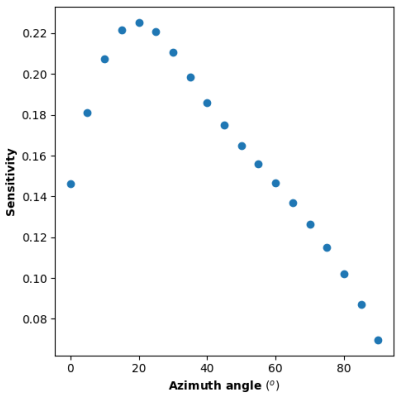


Fig. 4. Sensitivity plot for front-end module2 of FinFET.

Module	Parameter	Azimuth (°)	Sensitivity		Enhancement (%)
			(As-Is)	(To-Be)	
Front-end 1	Fin width	80	0.071	0.095	33.80
Front-end 2	Gate width	20	0.175	0.225	28.57

Table. 1. Sensitivity enhancement at the optimal azimuth angle for two FinFET front-end modules

Strong quantization of current-carrying electron states in δ-layer systems

Denis Mamaluy and Juan P. Mendez
Sandia National Laboratories
1515 Eubank SE, Albuquerque, NM 87123
email: dnmamal@sandia.gov

Introduction

Highly conductive δ-layer systems, i.e. thin, high-density layers of dopants in semiconductors are actively used as a platform for exploration of the future quantum and classical computing when patterned in plane with atomic precision. Such structures, with the dopant densities above to the solid solubility limit have been shown to possess very high current densities and thus have a strong potential for beyond-Moore and quantum computing applications. However, at the scale important for these applications, i.e. devices with sub-20 nm physical gate/channel lengths and/or sub-20 nm widths, that could compete with the future CMOS, the conductive properties of such systems are expected to exhibit a strong influence of size quantization effects.

Recently it has been demonstrated in [1] that to accurately extract the conductive properties of highly-conductive, highly-confined systems, an open-system quantum-mechanical analysis is necessary. Such open-system treatment, that can be conducted for instance using the Non-Equilibrium Green's Function (NEGF) formalism.

In this work we employ an efficient computational open-system quantum-mechanical treatment to explore the conductive properties of P δ-layer systems in Si (Si: P δ-layer), and the size quantization effects for sub-20 nm width devices. For devices widths W<10 nm, quantization effects are strong and it is shown that the number of propagating modes determines not only the conductivity, but the distinctive spatial distribution of the current-carrying electron states. For W>10 nm, the quantization effects practically vanish and the conductivity tends to the infinitely-wide device values.

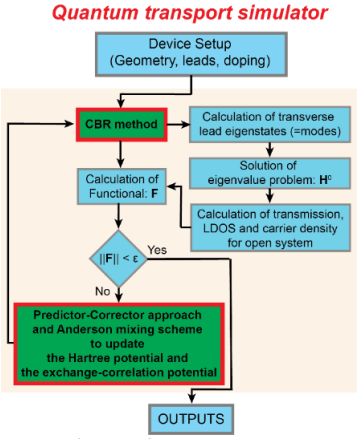


Fig. 1. Flow chart of the self-consistent quantum transport (QT) method [5,6,7,8,9].

Quantum Transport Framework

Our open-system QT framework [5,6,7,8,9] relies on a self-consistent solution of the Poisson-open system Schrödinger equation in the effective mass approximation and Non-Equilibrium Green's Function (NEGF) formalism.

Within the NEGF formalism, the current $J_{\lambda\lambda'}$ from lead λ to λ' is computed from the Landauer formula

$$J_{\lambda\lambda'} = \frac{2e}{h}\int T_{\lambda\lambda'}(E)(f_\lambda(E) - f_{\lambda'}(E))dE, \qquad (1)$$

where e is the electron charge, h is the Planck's constant, E is the energy, $f_\lambda(E) = f(E - E_F - qV_\lambda)$ is the Fermi-Dirac function within the leads, V_λ is the applied voltage to the lead, E_F is the Fermi level and $T_{\lambda\lambda'}$ is the electronic transmission from λ to λ'. The transmission function is given by

$$T_{\lambda\lambda'}(E) = Tr(\Gamma_\lambda \mathbf{G}_D \Gamma_{\lambda'} \mathbf{G}_D^\dagger), \qquad (2)$$

where $\Gamma_\lambda = i(\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^\dagger)$ are the coupling ($N_D \times N_D$)-matrices between the device and the leads, and $\mathbf{G}_D$ and $\mathbf{G}_D^\dagger$ are the retarded and advanced Green's functions ($N_D \times N_D$)-matrices of the coupled device with the leads (open-system device). The retarded Green's function matrix can be computed using the Dyson equation

$$\mathbf{G}_D = [\mathbf{I} - \mathbf{G}_D^0 \mathbf{\Sigma}]^{-1} \mathbf{G}_D^0 \qquad (4)$$

where $\mathbf{G}_D^0 = \sum_\alpha \frac{|\Phi_\alpha\rangle\langle\Phi_\alpha|}{E^+ - E_\alpha}$ and $\mathbf{\Sigma} = \sum_{\lambda=\lambda_1}^{\lambda_L} \mathbf{\Sigma}_\lambda$. The electron density matrix is defined as

$$\rho(r_i) = \sum_\lambda \int_{-\infty}^{\infty} \rho_\lambda(r_i, E) f_\lambda(E) dE, \qquad (5)$$

where $\rho_\lambda(r_i, E) = \frac{1}{2\pi} \mathbf{G}_D \Gamma_\lambda \mathbf{G}_D^\dagger$. We can notice that all matrices in the above matrix operations are of size ($N_D \times N_D$), where N_D is the total grid-points of the discretized device domain. Thus, for instance, the inversion matrix cost in Eq. 4 is of $O(N_D^3)$, and the calculation cost of the eigenstates of $\mathbf{H}_D^0$ in Eq. 4 is of $O(N_D N_e^2)$, where N_e is the number of calculated eigenstates.

To reduce the computational cost of these intensive calculations, we utilize the Contact Block Reduction (CBR) method. The CBR is an efficient method to calculate the electronic transmission function of an arbitrarily shaped, multi-terminal open device. Within this method, the N_D grid-points are subdivided into N_C boundary grid-points with the leads and N_{Di} interior grid-points of the device domain ($N_D = N_C + N_{Di}$, $N_{Di} >> N_C$). With this domain discretization, the self-energy matrix $\mathbf{\Sigma}$, the open-system device Hamiltonian $\mathbf{H}_D$, and the Green's function matrix of the open-system device $\mathbf{G}_D$ can be expressed as submatrices

$$\mathbf{H}_D = \begin{pmatrix} \mathbf{H}_C & \mathbf{H}_{CDi} \\ \mathbf{H}_{DiC} & \mathbf{H}_{Di} \end{pmatrix}, \mathbf{G}_D = \begin{pmatrix} \mathbf{G}_C & \mathbf{G}_{CDi} \\ \mathbf{G}_{DiC} & \mathbf{G}_{Di} \end{pmatrix} \text{ and } \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_C & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where the size of the sub-matrices $\mathbf{H}_C$, $\mathbf{G}_C$, and $\mathbf{\Sigma}_C$ are ($N_C \times N_C$), the size of the sub-matrices $\mathbf{H}_{CDi}$, $\mathbf{G}_{CDi}$ and $\mathbf{\Sigma}_{CDi}$ are ($N_C \times N_{Di}$), and the size of the submatrices $\mathbf{H}_{Di}$, $\mathbf{G}_{Di}$ and $\mathbf{\Sigma}_{Di}$ are ($N_{Di} \times N_{Di}$). After some algebra, the electronic transmission from lead λ to λ' can be computed as

$$T_{\lambda\lambda'}(E) = \mathrm{Tr}\left(\Gamma_{C_\lambda}\mathbf{G}_C\Gamma_{C_{\lambda'}}\mathbf{G}_C^\dagger\right), \quad (6)$$

where $\mathbf{G}_C = [\mathbf{I} - \mathbf{G}_C^0\mathbf{\Sigma}_C]^{-1}\mathbf{G}_C^0$ and $\Gamma_{C_\lambda} = i(\mathbf{\Sigma}_{C_\lambda} - \mathbf{\Sigma}_{C_\lambda}^\dagger)$. Similarly, the electron density matrix can be computed as

$$\rho(r_i) = \int_{-\infty}^{\infty} \Xi(E) f_\lambda(E) dE, \quad (7)$$

where $\Xi(E) = \frac{1}{2\pi}\frac{\mathbf{B}_C^{-1}\Gamma_C\mathbf{B}_C^{-1\dagger}}{(E^+ - E_\alpha)(E^- - E_\alpha)}$ and $\mathbf{B}_C = \mathbf{1}_C - \mathbf{\Sigma}_C\mathbf{G}_C^0$. Importantly, note that all matrices are now of size ($N_C \times N_C$), where $N_{Di} >> N_C$. Therefore, we can reduce considerably the computational cost by employing the CBR method in the NEGF formalism.

A summary of the algorithm implemented in our QT simulator is shown in Fig. 1. For the self-consistent solution of the non-linear Poisson equation, we employed a combination of the predictor-corrector approach and the Anderson mixing scheme.

Results and discussion

We have applied our QT framework to investigate effects of size quantization in Si: P δ-layer systems on the conductivity. The device is shown in Fig. 2.

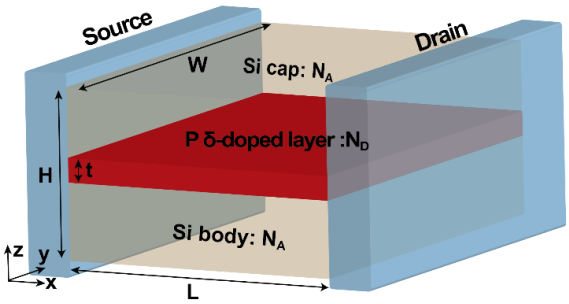


Fig. 2: The Si:P δ-layer device is composed of a Si body, a very high P-doped layer with an intrinsic gap, and a Si cap.

The corresponding dependence of the current on the device width (W) is shown in Fig. 3. The existence of the conduction steps due to each new propagating mode is well known experimentally since 1980's [10]. Here we report, however, that in highly-confined, highly-conductive δ-layer systems, the quantum number *m*, representing the number of propagating modes, determines not just the total current, but also the *spatial distribution of the corresponding current-carrying electrons* [11] as also shown in Fig. 3. The total number of propagating modes m depends on the number of peaks in the density of states (DOS) and is mainly determined by the three factors: 1) the δ-layer doping level N_D, 2) the the δ-layer doping thickness t and 3) the device width W. The spatial distribution of the current-carrying electron states, $n_{curr.-carr.}(y, z)$, can be obtained by performing the energy integration of the local density of electron states (LDOS) weighted by the corresponding current spectrum i(E) as: $n_{curr.-carr.}(y, z) = \int LDOS(y, z, E)i(E)dE / \int i(E)dE$. The spatial current-carrying electrons for the different modes is shown in Fig.3 a as insets in blue color, as well as the corresponding number of propagating modes. Additionally, the total electron density is also included in the figure as an inset in green color, demonstrating only weak spatial quantization along the y-direction. However, the specific portion of electrons with energies close to the Fermi level, i.e. the current-carrying states, do exhibit a strong spatial quantization. Indeed, for m = 1 the propagating mode reaches the maximum concentration at the center of the structure, the mode the corresponds to m = 2 is "excited" into the further penetration along the confinement direction (z-axis), leaving the center relatively depopulated (in terms of the current-carrying states), the mode m = 3 is again "pushed out" of the center along both z- and y- axis. One can further note that the modes m = 1, and m = 4, 6, etc. tend to form a regular "phase" distribution of the current-carrying states, while the modes m = 2, and m = 3, 5, etc. form "anti-phase" distributions that have the maximum current being carried in the different regions of space, separated by a few nanometers. When W → ∞, the number of propagation modesin y-direction becomes infinite m → ∞ as expected.

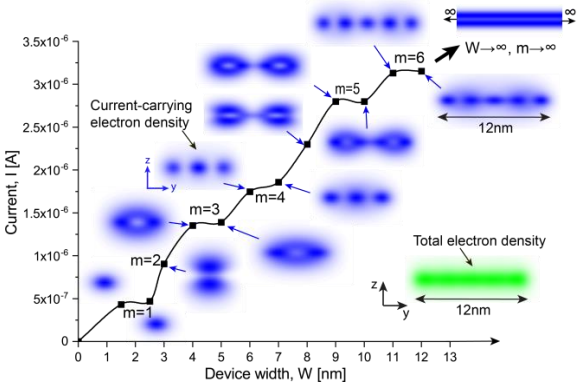


Fig. 3: Propagation modes for Si: P δ-layer systems. Current I vs device width W for δ-layer systems: the insets in blue color show the spatial distributions of current-carrying modes across a y-z plane, indicating the corresponding number of propagating modes m; Inset in green color shows the total electron density that includes all (not just current-carrying) occupied electron states for a device width of W=12nm. For all calculations, $N_D=1.0\times10^{14}cm^{-2}$, $N_A=5.0\times10^{17}cm^{-3}$, t=0.2nm and an applied voltage of 1mV.

Conclusions

We employed an efficient computational open-system quantum-mechanical treatment to explore the conductive properties of Si: P δ-layers, and the size quantization effects for sub-20 nm width δ-layers. We reported a strong spatial quantization of the current-carrying states for devices widths W<10 nm, which could be utilized in novel electronic δ-layer switches. The number of propagating modes could be controlled by external electric fields, thus strongly affecting the current. For devices widths W>10 nm, the quantization effects practically vanish and the conductivity tends to the infinitely-wide device values.

References

[1] D. Mamaluy, J. P. Mendez *et al.*,Commun Phys **4,**205(2021)
[2] D. Ward, *et al.*, EDFAAO 22, 4 (2020)
[3] S. Lee *at al*., Phys. Rev. B 84, 205309 (2011).
[4] D.J. Carter *et al.* Phys. Rev. B 79, 033204 (2009).
[5] D. Mamaluy *et al.*, J. Appl. Phys. 93, 4628–4633 (2003).
[6] D. Mamaluy *et al.*, Phys. Rev. B 71, 245321 (2005).
[7] H. R. Khan, *et al.*, IEEE T-ED. 54,784–796 (2007).
[8] D. Mamaluy, X. Gao, APL 106, 193503 (2015).
[9] X. Gao *et al.*, J. Appl. Phys. 115, 13707 (2014).
[10] B. J. van Wees et al., Phys. Rev. Lett. 60, 848 (1988).
[11] J. P. Mendez and D. Mamaluy, Commun Phys (submitted).

Abstract for SISPAD 2022 in Granada, Spain, 6-8 Sept 2022.

Non-local Transport Effects in Semiconductors Under Low-Field Conditions

M.G. Ancona*+ and S.J. Cooke+

*Department of Electrical and Computer Engineering, Florida State University
Tallahassee, FL 32310, mancona@fsu.edu
+Electronics Science and Technology Division, Naval Research Laboratory
Washington, DC 20375

Macroscopically non-local effects are common in electron transport in semiconductor devices, occurring whenever the mean free path (mfp) and/or the deBroglie wavelength are not small compared to geometry/flow length scales. When such effects are important, standard diffusion-drift (DD) theory becomes inaccurate and should be modified to account for the non-local effects. Density-gradient (DG) theory [1] is an example, and here we investigate another in which non-locality in the transport physics is significant.

The importance of non-local effects associated with the mfp is gauged by the Knudsen number defined as the ratio of the mfp to a characteristic length scale of the situation. When this number is appreciable, one is in the Knudsen regime and governing equations, which may be derived from the Boltzmann equation, are found to contain various gradient-dependent correction terms [2]. An analogous thing happens in a quantum mechanical context with DG theory wherein the equation of state for the electron gas contains an added gradient term that provides a lowest-order accounting for quantum non-locality. Here, using methods of classical field theory [3], we undertake a similar development but with the gradient corrections now capturing classical and/or quantum non-locality effects on scattering. The respective effects being analyzed are illustrated in Figs. 1b or 1c for flows involving surface scattering; Fig. 1a contrasts a parallel situation for an ordinary gas in the Knudsen regime.

In DD theory, the electron-lattice interaction is treated as local and instantaneous with the force felt by the gas being proportional to the average electron velocity and with the proportionality constant being the inverse mobility. The gradient correction must therefore also be in terms of velocity, and for reasons of rotational invariance we assume a Laplacian so that the material response function describing the electron-lattice interaction force is:

$$n\boldsymbol{E}^n(\boldsymbol{u}) = -\frac{n\boldsymbol{u}}{\mu_n} + \gamma_n n\nabla^2\boldsymbol{u} = -\frac{n\boldsymbol{u}}{\mu_n} + \gamma_n n[\nabla(\nabla\cdot\boldsymbol{u}) - \nabla\times\boldsymbol{\omega}] \qquad \boldsymbol{\omega}\equiv\nabla\times\boldsymbol{u}$$

where $\boldsymbol{\omega}$ is the vorticity, γ_n is a material parameter gauging the strength of the gradient effect in the gas, and the second equality follows from a vector calculus identity. Inserting this function into the force balance equation for the electron gas results in a corrected DD/DG theory which we refer to as generalized-gradient (GG) theory.

To illustrate GG theory (in the form of a corrected DD theory), we apply it to a long-channel Si FET under low-field conditions. We assume diffuse scattering at the Si-SiO_2 interface and that there is no vorticity injection from the contacts. Solution profiles obtained from the GG simulation with V_{DS} = 1V are plotted showing the electron density (Fig. 2a), the horizontal velocity (Fig. 2b), and the vorticity generated by the interface scattering (Fig. 2c). Profiles of density and velocity along a cutline across the boundary layer are shown in Fig. 3a. Lastly, IV curves are presented in Fig. 3b, comparing cases with full slip (specular scattering) and no slip (diffuse scattering). The former corresponds to the conventional treatment if it is scaled to match the latter by fitting the mobility and thereby defining a "channel mobility". These and other results, as well as full details on the theory, will be presented at the meeting.

Abstract for SISPAD 2022 in Granada, Spain, 6-8 Sept 2022.

Acknowledgement: The authors thank the Office of Naval Research for funding support.

1. M.G. Ancona, "Density-gradient theory: A macroscopic approach to quantum confinement and tunneling in semiconductor devices," *J. Comput. Electron.* **10**, 62-97 (2011).
2. D. Burnett, "The distribution of molecular velocities and the mean motion in a non-uniform gas," *Proc. Lond. Math. Soc.* **2**, 382-435 (1936).
3. C.A. Truesdell and R. Toupin, *The Classical Field Theories*, *Handbuch der Physik*, **III/1**, (Springer-Verlag, Berlin, 1960).

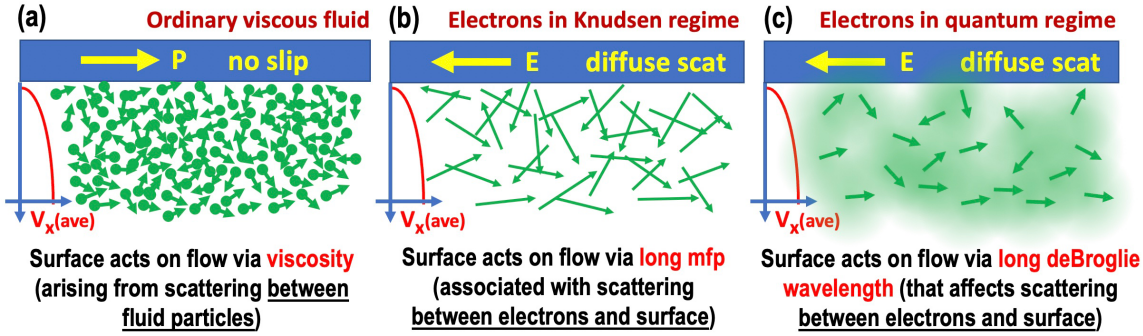


Figure 1. Depictions of the scattering and average velocity profiles for three flow situations near surfaces: (a) ordinary fluid mechanics; and electron flows with (b) a long mfp or (c) a long deBroglie wavelength.

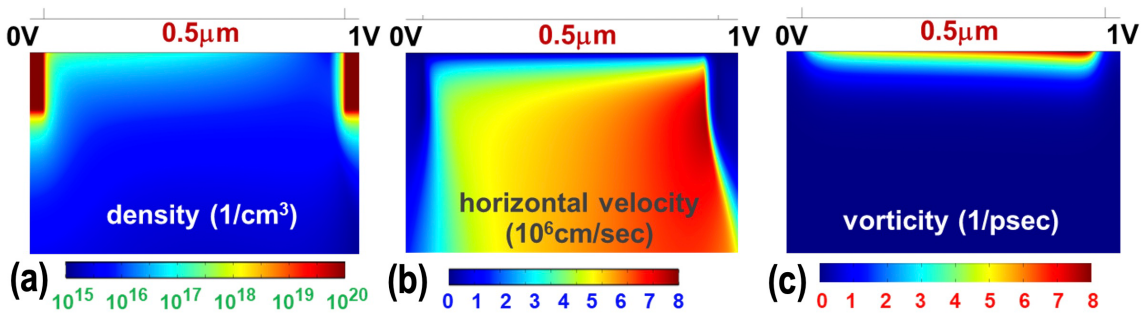


Figure 2. Field plots of (a) electron density, (b) horizontal velocity, and (c) vorticity inside a Si FET with V_{DS} = 1V as computed using GG theory.

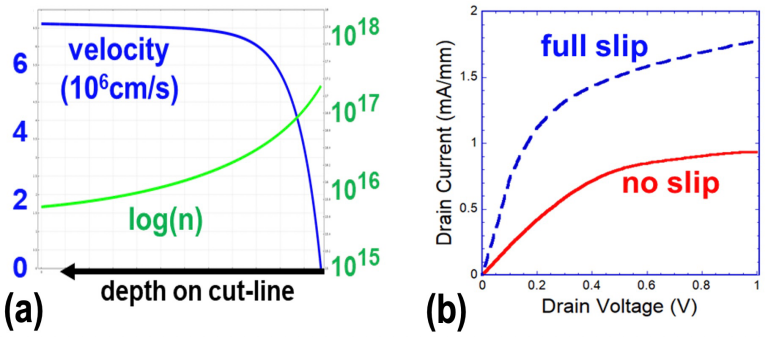


Figure 3. Plots of (a) cutlines across the channel in the FET of Fig. 2, and (b) the IV curves of the FET as computed assuming full slip (specular scattering) or no slip (diffuse scattering).

NANOACADEMIC TECHNOLOGIES
Coherent Modeling

info@nanoacademic.com
+1 (438) 387-4003
666 Rue Sherbrooke Ouest, Suite 802
Montréal, Québec, CANADA H3A 1E7
nanoacademic.com

Event: International Conference on Simulation of Semiconductor Processes and Devices (SISPAD 2022)
Presentation title: Tunneling leakage in ultrashort-channel MOSFETs—From atomistics to continuum modeling
Speaker: Dr. Raphaël J. Prentki, Research Scientist, Nanoacademic Technologies Inc.
Authors: Raphaël J. Prentki, Mohammed Harb, Chenyi Zhou, Pericles Philippopoulos, Félix Beaudoin, Vincent Michaud-Rioux, Hong Guo
(All authors are affiliated with Nanoacademic Technologies Inc., Montréal, QC, Canada)

Abstract

The channel lengths of some transistors are now nearing the nanometer [1], making these devices prone to direct source-to-drain tunneling (DSDT), a leakage mechanism commonly considered to set the end of Moore's law [2]. In a MOSFET, the probability for a charge carrier to undergo DSDT decays exponentially with channel length, source depletion length, and drain depletion length. Bound-charge engineering (BCE) is a recently introduced scheme where the depletion lengths of transistors can be controlled through effective doping by surface bound charges residing on the interface between a semiconductor and an adjacent oxide [3]. In this work, BCE is applied to reduce DSDT leakage current down to acceptable levels in MOSFETs with channels as short as 1.5 nm; the higher the oxide permittivity, the lower the DSDT leakage [Fig. 1 (a)–(c)]. As vehicles for this study, we consider n-type gate-all-around MOSFETs with (A) a 2-nm-wide silicon nanowire (NW) channel and (B) a silicon nanosheet (NS) channel [Fig. 1 (d)–(e)]. The silicon NW devices are modeled via state-of-the-art atomistic quantum transport simulations based on the nonequilibrium Green's function (NEGF) formalism and the tight-binding (TB) model, a simulation paradigm currently under development within our first-principles quantum transport software NanoDCAL+ [3, 4] which properly accounts for the atomistic effects, quantum confinement, and nonequilibrium quantum transport phenomena relevant in ultrascaled devices. On the other hand, the silicon NS devices, which are physically larger, are simulated within continuum models after calibration against NEGF–TB data. Specifically, we use the new NEGF–k·p feature under development within the QTCAD software suite (Quantum-Technology Computer-Aided Design), our finite-element-based quantum hardware modelling tool. It is found that short-channel device with high-permittivity oxides exhibit temperature-dependent DSDT, thereby maintaining their subthreshold swing at a value close to the thermal limit of 60 mV/dec at room temperature. BCE may thus pave a way toward ultrascaled MOSFETs. This work highlights the need for simulation platforms combining *ab initio* simulation tools, which can account for atomistic and quantum effects, with continuum models describing quantum transport at a mesoscopic scale, at which devices are too large to be described purely from first principles but small enough for tunneling to be significant.

1. S. B. Desai, S. R. Madhvapathy, A. B. Sachid, J. P. Llinas, Q. Wang, G. H. Ahn, G. Pitner, M. J. Kim, J. Bokor, C. Hu, H.-S. P. Wong, and A. Javey, “MoS2 transistors with 1-nanometer gate lengths,” Science, vol. 354, no. 6308, pp. 99–102, Oct. 2016. DOI: 10.1126/science.aah4698
2. J. Wang and M. Lundstrom, “Does source-to-drain tunneling limit the ultimate scaling of MOSFETs?” in International Electron Devices Meeting (IEDM) Technical Digest, pp. 707–710, Dec. 2002. DOI: 10.1109/IEDM.2002.1175936
3. R. J. Prentki, M. Harb, L. Liu, and H. Guo, “Nanowire transistors with bound-charge engineering,” Physical Review Letters, vol. 125, no. 24, p. 247704, Dec. 2020. DOI: 10.1103/PhysRevLett.125.247704
4. M. Harb, V. Michaud-Rioux, Y. Zhu, L. Liu, L. Zhang, and H. Guo, “Quantum transport modelling of silicon nanobeams using heterogeneous computing scheme,” Journal of Applied Physics, vol. 119, no. 12, p. 124304, Mar. 2016. DOI: 10.1063/1.4944649
5. F. Beaudoin, P. Philippopoulos, C. Zhou, I. Kriekouki, M. Pioro-Ladrière, H. Guo, P. Galy, “Robust technology computer-aided design of gated quantum dots at cryogenic temperature,” Applied Physics Letters, vol. 120, p. 264001, June 2022. DOI: 10.1063/5.0097202

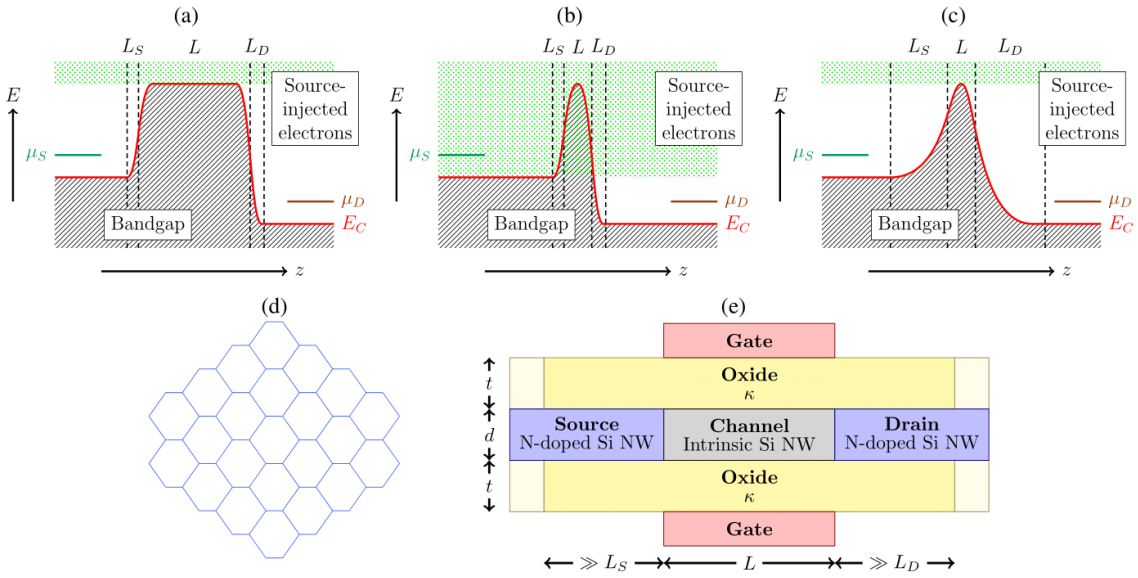


Figure 1: (a)–(c) Illustrations of off-state band diagrams (conduction band edge E_C as a function of position z along the transport direction) in the long-channel limit (a) and short-channel limit (b) of MOSFETs. Long-channel MOSFETs only exhibit thermionic emission, while short-channel MOSFETs exhibit both thermionic emission and direct source-to-drain tunneling (green-dotted region). (c) If the source and drain depletion lengths $L_{S,D}$ are sufficiently long, as may be achieved by BCE, DSDT can be significantly reduced. (d) Atomic structure of the unit cell of the silicon NW in the wireframe model. Lines represent interatomic bonds. The NW has a diameter of d = 2 nm and is grown in [110]. (e) Schematic of the n-type GAA silicon NW MOSFET investigated in this work. A cross section through the NW's axis of rotational symmetry is shown. An oxide with permittivity κ surrounds the channel as well as the source and drain depletion regions (with lengths $L_{S,D}$). By BCE, low-κ (high-κ) oxides lead to short (long) $L_{S,D}$. The nature of the oxide located far from these regions (light-yellow regions) does not affect DSDT.

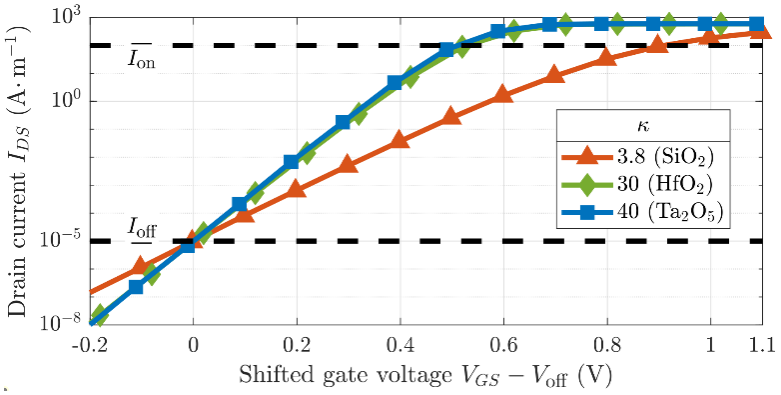


Figure 2: Transfer characteristics of ultrashort-channel GAA silicon NW MOSFETs [Fig. 1(e)] at a temperature of T = 300 K and a drain voltage of V_{DS} = 50 mV, as obtained from NEGF–TB simulations. Devices with various surrounding oxide permittivities κ are investigated. The devices have an extremely short channel length of L = 2.3 nm. The current is normalized by the perimeter of the NW's unit cell. The off-state voltage V_{off} is defined as the value of gate voltage V_{GS} at which the drain current I_{DS} is equal to I_{off} = 10^{-5} A/m. Due to increased tunneling leakage, the device with SiO_2 has a subthreshold swing of 130 mV/dec, to be compared to 75 mV/dec (72 mV/dev) for the device with HfO_2 (Ta_2O_5).

A Generalizable TCAD Framework for Silicon FinFET Spin Qubit Devices with Electrical Control

1st Qian Ding
Integrated Systems Lab
ETH Zurich
Zurich, Switzerland
dingq@iis.ee.ethz.ch

2nd Andreas V. Kuhlmann
Department of Physics
University of Basel
Basel, Switzerland
andreas.kuhlmann@unibas.ch

3rd Andreas Fuhrer
Quantum Technology & Computing
IBM Research Europe - Zurich
Zurich, Switzerland
afu@zurich.ibm.com

4th Andreas Schenk
Integrated Systems Lab
ETH Zurich
Zurich, Switzerland
schenk@iis.ee.ethz.ch

***Abstract*—We present a TCAD-based simulation framework established for quantum dot spin qubits with full-electrical control implemented on a silicon FinFET platform. It works down to 1 K and consists of a two-step simulation chain, from qubit initialization with DC bias to state manipulation using microwave signals. After calculating the microwave electric field response at the qubit locations, an average field polarization vector at each quantum dot is provided for further estimation of the Rabi coupling strength. We demonstrate the functionality via simulation of a recently reported two-qubit device in form of a 5-gate silicon FinFET. The framework is easily generalizable to future multi-qubit systems.**

I. Introduction

Scalability is vital for quantum computing, but a tough task from the aspect of physical implementations. One promising platform to overcome this challenge is given by quantum dot (QD) spin qubits embedded in multi-gate silicon FinFETs, which can be fabricated using standard CMOS technology. Recently, hole spin qubits hosted by double QDs in a 5-gate silicon FinFET that can operate above 4 K have been reported in [1] [2]. To scale up the system in the near future, a simulation-aided analysis for the design of full-electrical control is highly desirable, especially for the qubit state manipulation with microwave (MW) signals. For this purpose, we developed a TCAD-based framework operable down to 1 K which enables simulations of qubit state initialization with DC bias and manipulation by MW control signals. The MW response electric field (E-field) polarization vector averaged over each QD is extracted in post-processing steps, allowing for the further analysis such as Rabi frequency estimation. The cross-talk between multiple gates is also taken into account in the MW simulations via a 1st. order capacitive coupling model.

We illustrate the simulation framework by taking the example of the two-hole-qubit device reported in Ref. [1], but the generalization to multi-qubit devices (either electron or hole spin based) and the extension to include more functionalities like cross-talk between qubits and readout bus-line in a larger system are straightforward. The simulated device structure adapted from the fabricated one [1] is shown in Fig. 1. In the following, we first introduce the MW simulation method including the cross-talk estimation model. Then, MW simulation results for two qubits each hosted in a one-hole QD are presented to clarify how the average response field polarization vector at a given QD is extracted. Finally , as an application example, we demonstrate the impact of the numbers of holes in a QD on the response field polarization vector.

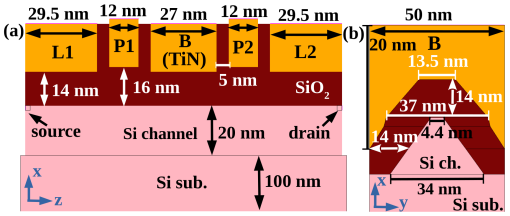


Fig. 1. Sketch of simulated 5-gate Si FinFET. (a) side view along fin direction, (b) cross section view perpendicular to fin direction under gate B.

II. Simulation Methodology

Modeling the MW field response in spin qubit devices has two special aspects. First, the charge distribution in the QDs is coupled self-consistently with the field response. Secondly, the device size (usually a few hundreds of nm) is much smaller than the wavelength of the applied MW signal (cm range). Therefore, commonly used techniques for MW simulation (like FDTD) are not proper options because of their high computational cost and the lack of self-consistency with the charge distribution. On the other hand, electrical small signal AC analysis fits well for this application, as long as the amplitude of the MW signal is much smaller than the applied DC bias. In S-Device, this technique proceeds by adding a small harmonic term to the original DC variable, i.e.

$$\xi_{total} = \xi_{DC} + \Delta\xi e^{iwt} \tag{1}$$

where ξ denotes potential (ϕ) or carrier density (n, p). The AC system containing the new variables is then solved in 1st-order of $\Delta\xi$ [3]. To calculate the response field, a two-step simulation is required. First, we run a quasi-stationary DC simulation (at T = 1 K) to generate QDs with a specific number of holes. The important effect of confinement in the fin tip is taken into account by the density gradient model [3]. Then, the electrical small-signal analysis is performed by ramping the AC voltage at gate P1. Since the MW field response that we are interested in is not a default output in S-Device, it needs to be extracted based on the following relations:

$$E = -\nabla\varphi \tag{2}$$

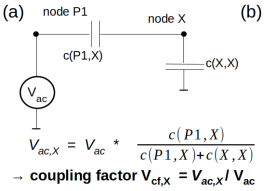
$$J_D = -i\omega\epsilon\nabla\varphi \tag{3}$$

$$\Re(E) = \Im(J_D)/\omega\epsilon \tag{4}$$

According to Eq. (4), the imaginary part of the displacement current response $\Im(J_D)$ (default output in S-Device)

is representative for the real E-field response $\Re(E)$. Their magnitudes differ only by a (computable) scaling factor, whereas the vector directions are exactly the same. This one-to-one correspondence facilitates the calculation of a response field polarization vector at each qubit location. Details will be discussed in Sec. III.

To take the cross talk between gates into account, we introduce a simplified capacitive coupling model based on the 1st-order estimation. This is achieved by running two-fold AC simulations. A first round is performed with the AC signal applied on gate P1 only, in order to extract the equivalent Y-matrix of the device. The obtained capacitance elements are then used to construct the simple voltage divider circuit model (see Fig. 2 (a)). The capacitive coupling factor $V_{cf,X}$ between P1 and any other gate (labeled as node X) can be calculated. Then, a second-round simulation is performed with AC signals also applied to all other gates, where their AC voltage amplitudes depend on the coupling factor $V_{cf,X}$. For example, Fig. 2 (b) shows the results of the coupling factors in case of one hole in each QD. The final field response is obtained after the second AC run including the gate cross talk.



(b) Gate coupling calculation for 1/1 hole @ QD1/2

Gate node X	cross cap. c(P1,X) [aF]	self cap. c(X,X) [aF]	$V_{cf,X}$ [1]
L1	11.050	15.707	**0.413**
B	10.882	26.920	**0.288**
P2	8.125e-4	24.237	3.352e-5
L2	5.341e-5	15.751	3.391e-6

Fig. 2. (a) Simplified 1st-order capacitive coupling circuit model for gate cross talk calculation. (b) Calculated voltage coupling factor for case of two one-hole QDs. The coupling is strong only for gates L1 and B that are close to the AC control gate P1 as expected.

III. Simulation Results

We first present the MW simulation results for a basic two-qubit case, with each qubit hosted at a one-hole QD. Then, we discuss the impact of the number of holes in QD1 on the field response.

A. Two-Qubit Case with One-Hole QDs

First, we run DC simulation for qubit state initialization, with gate biases set to values adapted from the experiment. The obtained hole density profile (see Fig. 3 (a)) shows the formation of two one-hole QDs. The resultant real part of the potential response is plotted in Fig. 3 (b), where its gradient relates to the response field we are interested in. Since this response is the lowest in the region with the highest hole density, a response singularity appears at the center of each QD (see regions labeled by white dashed lines in Fig. 3 (c)). The presence of such singularities makes it hard to directly assign a single field vector to a particular QD. Thus, we introduce a normalized field vector averaged over the QD volume. It is calculated by integrating each component of the field vector over the dot volume and dividing by the root sum square (i.e. normalizing by the volume integral of the field magnitude). The results are shown in Fig. 3 (d).

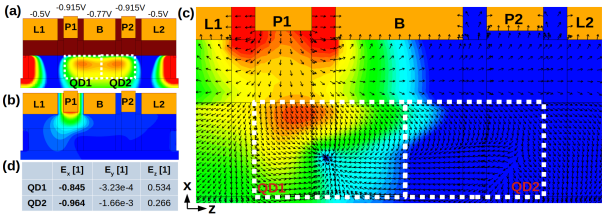


Fig. 3. (a) DC hole density profile in fin direction with 1/1 hole at QD1/2. Applied DC biases are labeled on top of the corresponding gates. White dashed lines indicate the QD volume used for integration to obtain the number of holes. Corresponding (b) normalized real part of the potential and (c) imaginary part of the displacement current response, taken at V_{ac} = -12 mV. (d) Calculated components of the averaged response field vector at QDs. Color bar ranges from 1e14 to 1e18 cm^{-3} in (a), 0 (blue) to 1 (red) in (b), and from 1e2 (blue) to 1e4 (red) $Acm^{-2}V^{-1}$ in (c).

B. Impact of Number of Holes at QDs

Using the method above, we study the impact of the number of holes at QD1 on the response field, as this number can be hard to determine experimentally. Results for two cases, 3/3 and 5/3 holes at QD1/2, are shown in Fig. 4. Although the dominant field polarization in both QDs remains unchanged with more holes in QD1, the field vector at QD1/2 tends to have a stronger x/z-component and a weaker z/x-component. For QD1, this is because the increased charge there distributes the field response more homogeneously along z (see colormap comparison under P1). Consequently, the z-component of the field contributes less over the integrated dot volume. For QD2, the slightly larger z-component comes from the increased electrostatic impact along z due to more holes at QD1.

(a) 3/3 holes @ QD1/2

	E_x [1]	E_y [1]	E_z [1]
QD1	**-0.956**	-3.97e-4	0.292
QD2	**-0.836**	-1.67e-3	0.549

(b) 5/3 holes @ QD1/2

	E_x [1]	E_y [1]	E_z [1]
QD1	**-0.984**	-3.12e-4	0.180
QD2	**-0.827**	-1.69e-3	0.563

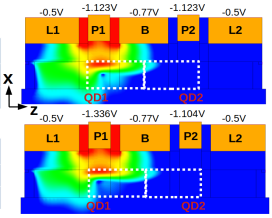


Fig. 4. Simulated field response for case (a) 3/3 and (b) 5/3 holes in QD1/2. Tables contain the components of the average response field vector. Field response magnitudes are plotted on the right. White dashed lines indicate the dot volume used for integration. Color bar range is kept the same as in Fig. 3 (c). Numbers on top of the gates indicate the corresponding DC biases to obtain 3/3 and 5/3 holes under QD1/2.

IV. Conclusion

A TCAD simulation framework developed for future large-scale silicon FinFET spin qubit devices allows to calculate the average MW E-field response vector at each QD. A reported two-qubit device is analyzed, showing that the MW response field polarization is very sensitive to the number of holes in the QD under the AC gate used for qubit control.

References

[1] S. Geyer, L. C. Camenzind, L. Czornomaz, V. Deshpande, A. Fuhrer, R. J. Warburton, D. M. Zumbühl, and A. V. Kuhlmann, "Self-aligned gates for scalable silicon quantum computing," *Applied Physics Letters*, vol. 118, no. 10, p. 104004, 2021.

[2] L. C. Camenzind, S. Geyer, A. Fuhrer, R. J. Warburton, D. M. Zumbühl, and A. V. Kuhlmann, "A hole spin qubit in a fin field-effect transistor above 4 kelvin," *Nature Electronics*, Mar. 2022.

[3] Synopsys, *Sentaurus Device User Guide*, Version R-2020.09.

A Simulation Methodology for Superconducting Qubit Readout Fidelity

Hiu Yung Wong*, Yaniv Jacob Rosen+, Kristin Beck+, Prabjot Dhillon, and Jonathan L Dubois+
Electrical Engineering, San Jose State University, CA, USA
+Lawrence Livermore National Laboratory, Livermore, CA, USA
*hiuyung.wong@sjsu.edu

Abstract – Motivation and Achievement

Superconducting qubits are one of the most promising quantum computing architectures [1]. While a qubit needs to have enough isolation to achieve a long coherence time, it should also be allowed to interact with the outside world for the readout operation. Often, a resonator is coupled to a qubit to allow dispersive readout, in which the resonator will experience a resonant frequency shift depending on the final state of the qubit [2]. This frequency shift is called the Cross-Kerr, χ. The larger the χ, the easier it is to distinguish the qubit $|0\rangle$ and $|1\rangle$ states. However, this will also result in a shorter coherence time. The distinguishability of the $|0\rangle$ and $|1\rangle$ states also depends on the readout pulse power and duration, the resonator scattering matrix, and the noise from the circuits. Therefore, it is important to co-optimize the resonator design, qubit-resonator coupling, and reading pulse length and power with the noise taken into account.

In this work, a simulation methodology for superconducting qubit readout fidelity is proposed and implemented to allow the aforementioned co-optimization. Parameters are taken from an actual physical qubit system. *The simulation is calibrated to the experiment*. This can then be used to guide the design and optimization of a superconducting qubit readout system. *As an example, it is found that the system can still maintain > 99% fidelity even if the input power is reduced by 10dB or if the readout pulse width is halved.*

Qubit Readout System

Fig. 1 shows the hardware system used in this paper. Quantum Machine OPX is used as the control hardware [3]. A -47dBm 3.5μs (t_p) reading pulse of 7.246245GHz is used. After three attenuation stages (-60dB in total), the pulse reaches the input port (port 1) of the resonator coupled to a qubit at 10mK. The qubit is tantalum-based with a high coherence time (>0.3ms) [4]. The signal from the output port (port 2) of the resonator is then amplified by a Traveling Wave Parametric Amplifier (TWPA) (+20dB) at 10mK, a HEMT amplifier at 4K (+40dB), and a 300K amplifier (+40dB). Quadrature measurement is performed on the amplified output signal, which represents the S_{21} of the resonator/qubit system, to distinguish the qubit $|0\rangle$ and $|1\rangle$ states. The χ of the system is 114kHz.

Simulation Methodology

Fig. 2 shows the simulation framework. The framework uses Ansys HFSS [5] to perform the scattering matrix simulation of the resonator. The S_{21} obtained is then fed into a MATLAB program to simulate the readout process. White noise corresponding to the amplifiers is generated in the frequency domain. There are three major noise sources. The first one is the quantum noise due to the photon number fluctuation after the resonator. The second one is the noise due to the TWPA. Since TWPA is a quantum-limited amplifier, therefore, at the best case, it only reduces the signal-to-noise ratio by half [6]. This is equivalent to adding 3dB of noise to its output. Thirdly, the two low noise amplifiers contribute thermal noise equivalent to T_{eff} = 1.5K [7] and T_{eff} = 54K [8], respectively, with a noise spectral density of $4kT_{eff}R$, where k is the Boltzmann constant and R is 50Ω. The output pulse from the resonator is simulated by multiplying the attenuated input pulse and the S_{21}. The total white noise is then added to the output pulse in the frequency domain. The real and imaginary parts at the reading frequency are extracted to simulate the quadrature measurement. 1000 random runs are performed to obtain the statistics.

Simulation Results

The resonators are designed to have eigenfrequencies of 7.252456GHz and 7.252612GHz, to emulate the coupled qubit's $|0\rangle$ and $|1\rangle$ states, respectively. This gives a χ of 156kHz, which is similar to that of the hardware. Fig. 3

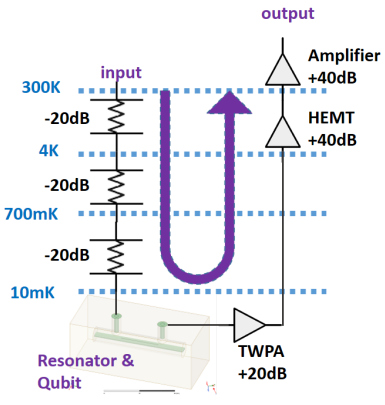


Figure 1: The qubit system used. The readout path is highlighted.

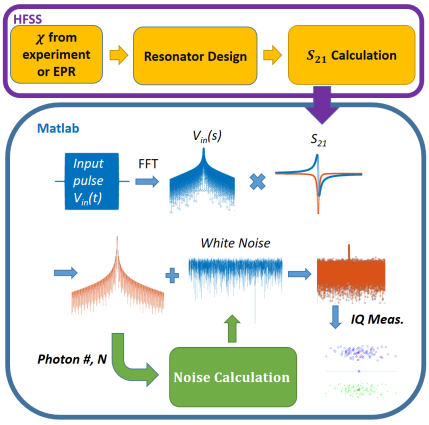


Figure 2: Ilustration of the simulation flow.

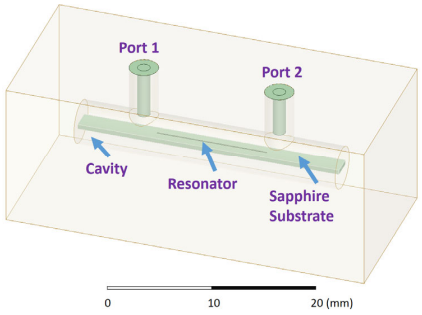


Figure 3: The cavity and resonator used in the HFSS simulation.

shows the design of the cavity and the resonator with Q ~ 48k, similar to the experimental value. If experimental χ is not available, it can be obtained using the Energy Participation Ratio (EPR) method with HFSS [2]. The reading pulse frequency is $f = 7.252534$ GHz , which is the average of the two resonator frequencies. Based on the simulation, the number of photons entering port 1 is about 14k and $N = 3746$ photons are emitted from port 2. Therefore, the quantum noise power is $hf\sqrt{N}/t_p$, where h is the Planck's constant. The total noise power is found to be -27.7dBm. Fig. 4 shows the output signal before and after the chain of amplifiers for resonator coupled with qubit with states $|0\rangle$ and $1\rangle$. It can be seen that the noise reduces the distinguishability. Fig. 5 shows the fidelity of the qubit based on experimental quadrature measurement and simulation which match each other pretty well. This framework is then used to study how the input pulse power will change the fidelity of the system. The simulation shows that it can maintain >99% fidelity even if the input power is reduced by 10dB or if the readout pulse is 1.5μs. Therefore, there is still room to further optimize the system (Fig. 6).

Conclusions

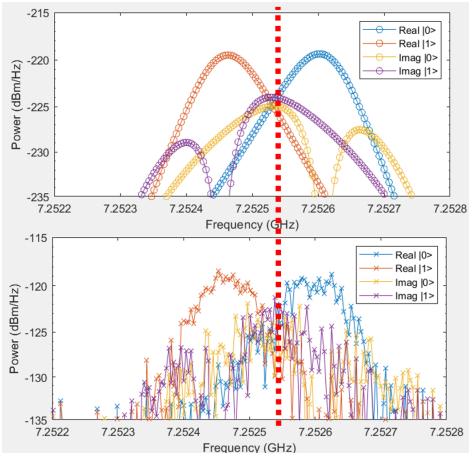


Figure 4: The real and imaginary compoents of the signal after the resonantor before adding the quantum noise (Top) and after the amplification chain in Fig. 1 (Bottom). The red dotted line indicates the reading pulse frequency.

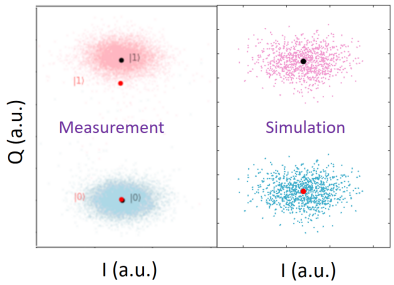


Figure 5: The quadrature measurement (Left) and simulation (Right) for reading $|0\rangle$ and $1\rangle$ states.

In this paper, a simulation method for predicting superconducting qubit fidelity is proposed and implemented. It is verified against hardware data. The simulation result matches the experimental result well after calibration. It is then used to predict the fidelity of the system with reduced reading pulse power. It is found that the fidelity is still > 99% even with a 10dB reduction in the reading pulse power or if the readout pulse width is halved. This can be used for the further optimization of the system.

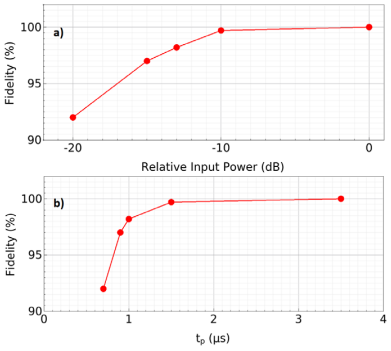


Figure 6: Simulated fidelity of the qubit readout system as a function of a) relative input pulse power (relative to -47dBm) and b) readout pulse duration.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 2125906. The authors thank MITLL and IARPA for allowing us to use their TWPAs. Prepared in part by LLNL under Contract DE-AC52-07NA27344.

References

[1] F. Arute et al., 'Quantum supremacy using a programmable superconducting processor," Nature 574, 505–510 (2019). https://doi.org/10.1038/s41586-019-1666-5.

[2] Z. K. Minev et al., " Energy-participation quantization of Josephson circuits," npj Quantum Inf 7, 131 (2021). https://doi.org/10.1038/s41534-021-00461-8

[3] https://www.quantum-machines.co/opx+/

[4] Alexander Place, et al., "New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds,"Nat. Comm 12, 1779 (2021), https://doi.org/10.1038/s41467-021-22030-5.

[5] Ansys® Academic Research HF, Release 2021 R1

[6] https://www.ibm.com/blogs/research/2020/01/quantum-limited-amplifiers/

[7] https://www.lownoisefactory.com/products/cryogenic/lnc-4-8-ghz/

[8] https://www.lownoisefactory.com/products/roomtemp/1-15-ghz/

On the noise-sensitivity of 2-qubit entangling gates implemented with a silicon quantum dot system

Hoon Ryu
Korea Institute of Science and Technology Information, Daejeon 34141, Republic of Korea
Electronic mail: elec1020@kisti.re.kr

Introduction

Electron spins in Silicon (Si) are promising for designs of universal quantum gates due to their extremely long coherence time [1]. The electrode-driven Si quantum dot systems have been extensively studied to realize electron spin qubits as they, in principle, can be fabricated with industry-standard processes. Recently, a controlled-NOT (CNOT) logic, whose gating can be controlled with a single microwave pulse, has been realized with a Si double quantum dot (DQD) system [2] and the theoretical background [3]. The correlation between CNOT-fidelity and charge noise, however, is not yet clear even though charge noise is omnipresent in Si devices. Here, we computationally explore how charge noise affects entangling logics implemented with a Si DQD structure including the reported single-step CNOT operation [2].

Methods

The physically realized DQD system is represented in a 2D manner as shown in Figure 1(a), since the reported structure is quite long along the [001] (Z) direction [2]. The bias-dependent electrostatics are simulated with bulk physics augmented with electronic structure calculations based on the effective mass model [4]. A DC magnetic field along the Z-direction (B_Z), which is generated with a horseshoe-shaped micro-magnet in the real experiment, is incorporated into simulations with a spatial distribution driven by Neumann *et al.* [5], and the time responses of spin qubits are calculated with the two-spin Heisenberg Hamiltonian [2] whose matrix elements are determined with DQD electrostatics and time-varying control pulses.

Results and Discussion

DQD Initialization: The left subfigure of Figure 1(b) shows the charge stability diagram that is simulated at a barrier gate bias (V_B) of 200mV and a middle gate bias (V_M) of 400mV, where two numbers in each parenthesis represent electron populations of two QDs. The starting step of qubit operations is to initialize the DQD system such that the lowest down-spin ($|\downarrow\rangle$) state of each QD is filled with a single electron. Here we initialize the system to a 2-qubit $|\downarrow\downarrow\rangle$ state (= $|\downarrow\rangle\otimes|\downarrow\rangle$) by setting a left ($V_L$) and a right gate ($V_R$) bias to 540mV and 570mV, respectively, since it is known that *symmetric-biasing* (*i.e.*, potential shapes of two QDs become identical) is beneficial for reducing the sensitivity of qubit interactions

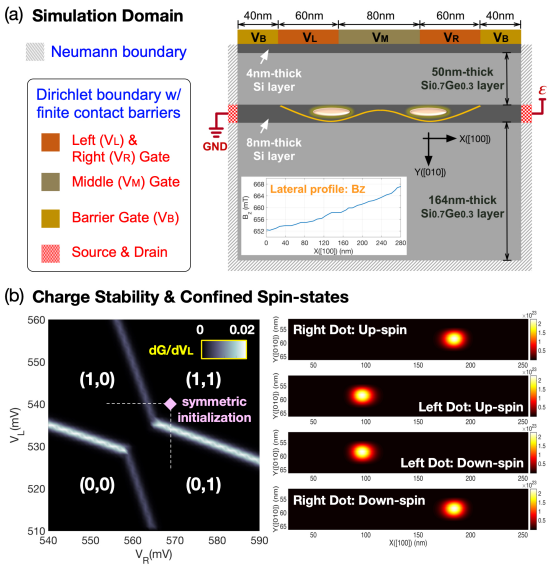


Fig. 1. (a) A 2D domain representing a Si DQD system that is assumed be infinitely long along the [001] direction. Quantum confinement along the [100] direction is formed with biases imposed on top electrodes. Confinement along [010] direction is created by the band offset between Si and SiGe layers. (b) A charge stability calculated as a function of V_L and V_R (V_M = 400mV, V_B = 200mV), and electron density profile at lowest $|\uparrow\rangle$ and $|\downarrow\rangle$ states of two QDs. At (V_L, V_R) = (540mV, 570mV), QDs are symmetrically initialized, and their Zeeman-splitting energies become 18.45GHz (right) and 18.31GHz (left).

to charge noise [6]. The right subfigure of Figure 1(b) shows the spatial distribution of electrons at $|\uparrow\rangle$ (lowest up-spin) and $|\downarrow\rangle$ state of two QDs. The Zeeman-splitting energy in the right (E_{ZR}) and left QD (E_{ZL}) becomes 18.45GHz and 18.31GHz due to the inhomogeneous B_Z profile along the [100] (X) direction (see Figure 1(a)).

2-qubit entangling operation: Figure 2(a) shows how E_{ZL}, E_{ZR}, and exchange interaction between left and right $|\downarrow\rangle$ state (J) changes with increasing V_M (at V_L = 540mV, V_R = 570mV). Varying V_M in a 10mV range (affecting the barrier height between QDs) does not drive remarkable changes of E_{ZL} and E_{ZR}. The sensitivity of J to V_M, however, is extremely large so J at V_M = 400mV and 408.1mV is calculated as 75.6KHz and 19.3MHz, respectively. At V_M = 400mV where J is still in the order of KHz, we have shown that both QDs can be

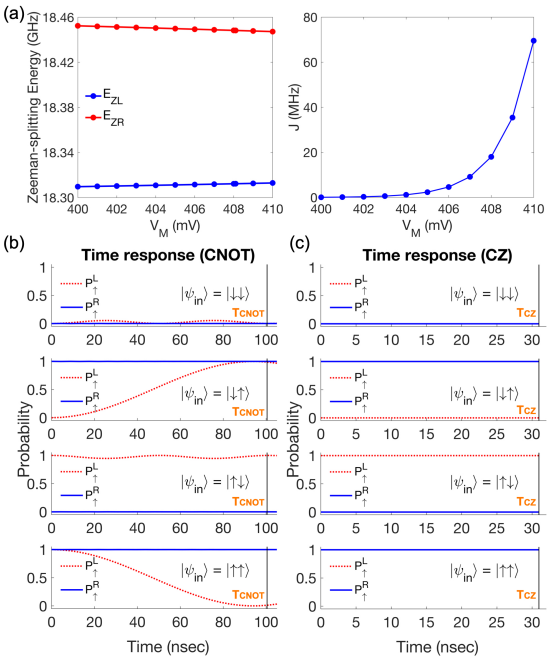


Fig. 2. (a) Interaction between inter-QD spin states is quite sensitive to V_M, so ΔV_M of a few mV changes J by an order of magnitude. The Zeeman-splitting energy of electron spin in each QD, however, rarely depends on V_M, so J can be controlled almost independently of E_{ZL} and E_{ZR}. (b) 2-qubit time response obtained at J = 19.3MHz (V_M = ∼408mV) with a [010]-oriented AC magnetic pulse and (c) with no AC magnetic pulses. The CNOT logic is secured at 100.4nsec (T_{CNOT}) while the CZ logic can be achieved at 32nsec (T_{CZ}) under only [001]-oriented DC field that is generated from a micro-magnet.

individually addressed [4]. When J reaches 19.3MHz at V_M = 408.1mV, the interaction is not negligible and we have possibility for implementation of entangling logics. Figure 2(b) shows the 2-qubit time response calculated against the 4 input states ($|\downarrow\downarrow\rangle$, $|\downarrow\uparrow\rangle$, $|\uparrow\downarrow\rangle$, $|\uparrow\uparrow\rangle$). Due to the non-negligible J, the resonance frequency of the right QD depends on the occupied spin state of the left QD, and the CNOT logic can be secured at 100.4nsec with a single [010]-oriented AC magnetic pulse ($B_Y(t)$), where $B_Y(t)$ (an amplitude of 4.98MHz and a driving frequency of 1.83GHz) is incorporated into the Heisenberg Hamiltonian for modeling of the response.

Another important entangling logic for universal quantum computing, controlled-Z (CZ) operation, can be also implemented with the same DQD structure. Unlike the 1-step CNOT case, this does not even require additional time-varying magnetic pulses. Figure 2(c) shows the 2-qubit time response that is simulated with only above-mentioned DC control factors (DC biases imposed on top electrodes and B_Z profile). Here, we see that the CZ logic is secured at ∼32nsec, which is much faster than the the CNOT case since the operation is solely determined with the DC field (B_Z) that is much larger in magnitude

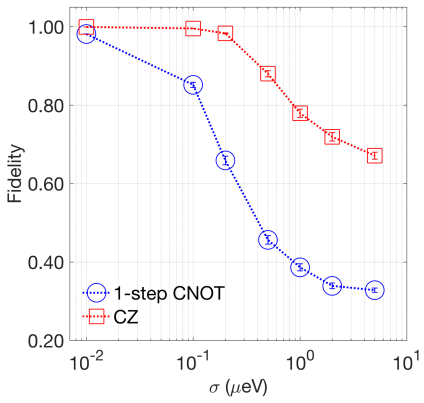


Fig. 3. In a noise-free condition, the fidelity of CNOT and CZ logic is 98.34% and 99.94%, respectively. However, the fidelity reduces once we incorporate charge noise into simulations, which is done by adding randomly generated noisy potential profiles to noise-free solutions. The noisy value is generated per grid in the DQD domain under a zero-mean gaussian distribution of standard deviation σ. Results of 1,000 simulations show CZ is much more robust to charge noise than CNOT.

than the AC pulse ($B_Y(t)$) used to get the CNOT logic.

Sensitivity to charge noise: In a noise-free condition, the fidelity of CNOT and CZ logic becomes 98.34% and 99.94%, respectively. Once we start to incorporate charge noise into simulations by disturbing the DQD potential distribution with random values generated under a zero-mean gaussian distribution of standard deviation σ, the fidelity is reduced in both cases. But, as shown in Figure 3, we recognize that the CZ logic is much more robust to noise than the CNOT case so its fidelity remains around 70% even at σ = 5μeV (∼32% for CNOT).

Conclusion

A preliminary modeling study on the noise-robustness of entangling operations in Si DQD systems is presented.

Acknowledgments

This work has been supported by the Korea Institute of Science and Technology Information (KISTI) institutional R&D program (K-22-L02-C04), The NURION supercomputer has been extensively used.

References

[1] J. Muhonen *et al.*, "Storing quantum information for 30 seconds in a nanoelectronic device." *Nat. Nanotechnol.* 9, 986 (2014).
[2] D. Jajac *et al.*, "Resonantly driven CNOT gate for electron spins." *Science* 359, 439 (2018).
[3] M. Russ *et al.*, "High-fidelity quantum gates in Si/SiGe double quantum dots." *Phys. Rev. B* 97, 085421 (2018).
[4] J. Kang *et al.*, "Exploring the behaviors of electrode-driven Si quantum dot systems: from charge control to qubit operations," *Nanoscale* 13, 332 (2021).
[5] R. Neumann *et al.*, "Simulation of micro-magnet stray-field dynamics for spin qubit manipulation," *J. Appl. Phys.* 117, 193903 (2015).
[6] M. Reed *et al.*, "Reduced Sensitivity to Charge Noise in Semiconductor Spin Qubits via Symmetric Operation," *Phys. Rev. Lett.* 116, 110402 (2016).

RF simulation platform of qubit control using FDSOI technology for quantum computing

H. Jacquinot[1], R. Maurand[2], G. Troncoso Fernandez Bada[2], B. Bertrand[1], M. Cassé[1], Y. M. Niquet[2], S. de Franceschi[2], T. Meunier[3], M. Vinet[1]
[1]Univ. Grenoble Alpes, CEA, LETI, F-38000 Grenoble, France, E-mail: helene.jacquinot@cea.fr
[2]Univ. Grenoble Alpes, CEA, IRIG/DEPHY, Grenoble, France
[3] CNRS Institut Néel, Grenoble, France

***Abstract*—In this paper, we report on simulations of an Electron Spin Resonance (ESR) RF control line for semiconductor electron spin qubits. The simulation includes both the ESR line characteristics (geometry and configuration, stack and material properties) and the electromagnetic (EM) environment at the vicinity of the qubits such as gates and interconnect network. With the accurate assessment of the magnetic and electric field distribution, we found that the EM environment of the qubits contributes significantly to the ESR line efficiency for spin control characterized by the magnetic over electric field ratio generated at the qubit location.**

***Index Terms*—ESR, electron spin qubits, Quantum computing**

I. Introduction

Thanks to their long coherence time and their compatibility with advanced semiconductor manufacturing, electron spin qubits are expected to bring breakthrough in Quantum computing technologies [1]–[3]. To enable fabrication of a multi-qubits demonstrator, spin control modules need to be developed together with the qubits full integration flows. Spin qubit control can be achieved by electron spin resonance (ESR) [4]. It consists in applying to the qubit a resonant AC magnetic field generated by the AC current flowing through an RF line at the vicinity of the qubit [5]–[7]. Usually, the ESR RF line is simulated without considerating the electromagnetic (EM) effects of the surroundings of the qubit such as interconnects, dummies and gate structure. Here the simulations aim at describing a realistic environment including the qubits and a real BEOL (Back-end-of-Line) process in a FDSOI technology operating at cryogenic temperature [8], [9].

II. Electromagnetic simulation platform

In this work, we use EM simulation deck for assessing specific ESR line figure-of-merits (FoM). Simulations are realized using HFSS from Ansys and CST from Dassault System, which are finite element EM solvers, for coping with both ESR and quantum dots (QDs) co-design and multi-scale requirements.

We will focus in this paper on the magnetic over electric field ratio (B/E) for a given ESR line input power, as the AC magnetic field is the one used for ESR (directly proportional to the spin Rabi frequency), whereas E field is the parasitic one, potentially heating the sample and leading to qubits improper operations [5], [6], [10]. In Fig. 1, we propose a classification of the ESR line FoM and we demonstrate in this paper that the B/E ratio is both ESR line geometry and configuration, stack and material and EM environment dependent. The simulated structure is described in Fig. 2 and 3.

A. ESR line configuration and stack impact

The ESR line geometry are either based on a coplanar stripline (CPS), a coplanar stripline using a balun or a coplanar waveguide (CPW) [6], terminated by a short-circuit placed near the Qubit QDs, where the ratio B/E is to be maximized. Figure 4 compares the FoM of the different ESR line configurations and points out the trade-off between maximal B field and maximal B/E ratio.

The impact of the stack for designing the ESR line is also studied considering two cases: either the fabrication of the CPS line at the gate level or at the first metallization level M1 (Fig. 4). The results are straightforward: the decrease of B/E when using M1 level is due to the decrease of B field as the distance to the QDs is increased.

B. EM environment and positioning impact

Simulation results show that taking into account all the conductive, dielectric layers and polysilicon dummies has a strong effect on the electric field, making it much more inhomogeneous along the line, contrary to the usually simulated simple ESR line geometry evaluation. For the ESR line represented in Fig. 2, the polysilicon dummies of few nanometers can reduce up to $75\,\%$ the E field thanks to their screening effect.

Inversely, interconnect network of the QDs can increase the electric field locally and degrade the B/E. When connecting the QDs with exchange gates in a face-to-face configuration [11], extra gate interconnect can lead to extra E field. Thus, a precise multi-scale description of the device and its EM environment in the simulation platform has to be added to the usual ESR stand-alone device evaluation for accurate FoM assessment, as summarized in Fig. 5.

III. Comparison with experimental data

A co-design 'ESR line/qubit' using a dedicated state-of-the-art CMOS FDSOI technology [11] has been fabricated and characterized at room and cryogenic temperatures using a Vector-Network-Analyzer with standard and on-chip calibration. For the 28FDSOI conductive layers, we have used their RRR (Residual-resistance ratio) values based on $4\,K$ experimental results and have adjusted the conductivity for the nano-antenna part (Fig. 2) in M1 to $3.10^7\,S/m$. Fig. 6 outlines the low resistive access line impact and the quasi-static behaviour of the nano-antenna (quasi-constant resistance over frequency) due to a very high wavelength to geometric length ratio. As shown in Fig. 6 and 7, CEA-LETI simulation platform gives very good agreement between measurements and simulations over a wide frequency range up to $20\,GHz$. Moreover, the obtained results in Fig. 7 show wideband and low-loss characteristics of the ESR line, with insertion loss parameter (S11) significantly reduced at cryogenic temperature mainly due to a higher M1 conductivity.

IV. CONCLUSION

Evaluation of the ESR line control EM fields with QDs is performed using a dedicated simulation platform. While only ESR line geometry impact had been studied up to now, we also include in this study the technological stack and the EM environment, considering dummies and interconnects in the vicinity of the QDs, and simulations results clearly indicate their important impact. Finally, this simulation platform being experimentally validated, it can be used as a predictive tool to co-design ESR line and QDs and to explore new materials like superconductors for control efficiency optimization.

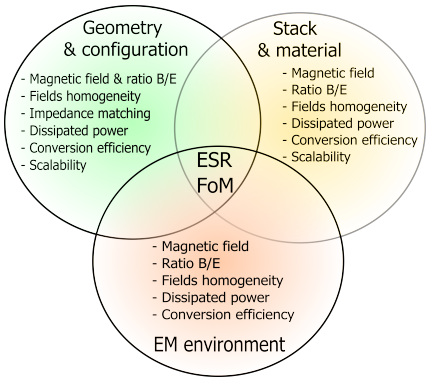


Fig. 1. Classification of ESR FoM: ESR line geometry and configuration, stack and material and EM environment dependent.

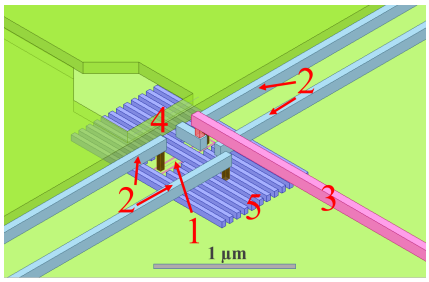


Fig. 2. Structure used for 3D electromagnetic (EM) simulations to assess RF performances for spin manipulation using Electron Spin Resonance (ESR) 1) silicon active, QD, 2) Gates and reservoirs, 3) Top gate, 4) ESR line nano-antenna in M1 level, 5) Dummies (polysilicon), illustration from [8].

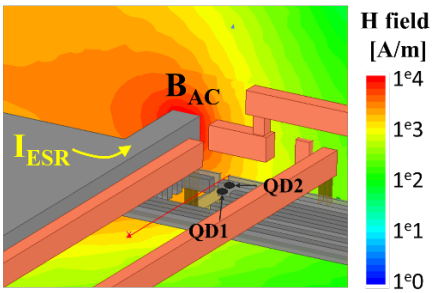


Fig. 3. Cross-sectional view in the Si QD2 plane of the magnetic field of the ESR line with two quantum dots at $10\,GHz$ with a $-7\,dBm$ input power, illustration from [8].

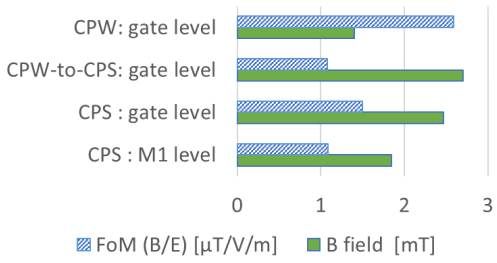


Fig. 4. ESR line configuration and stack impact on FoM, average fields at QDs locations, Pin=$-7\,dBm$.

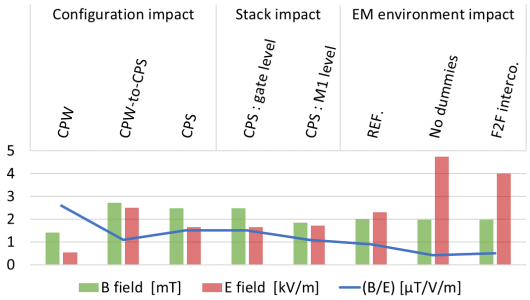


Fig. 5. Summary of impact on ESR line FoM, average fields at QDs locations, Pin=$-7\,dBm$ (REF.: CPW-to-CPS at M1 level, with dummies and with no F2F interconnect).

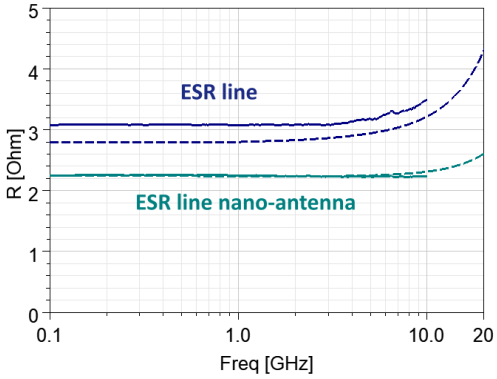


Fig. 6. Comparison of the ESR line overall resistance with the ESR nano-antenna resistance (extracted by de-embedding) at cryogenic temperature in the [$100\,MHz$ – $20\,GHz$]: experimental results (lines) and EM simulations (dashed lines).

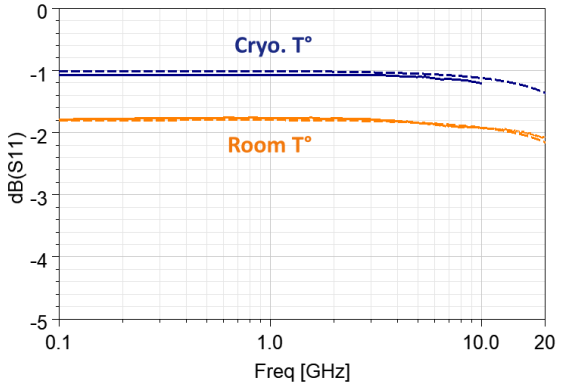


Fig. 7. Comparison of experimental results (lines) with EM simulations (dashed lines) S parameters in the [$100\,MHz$ – $20\,GHz$] of a co-design 'ESR line/qubit' using a double quantum dots with exchange gate in a state-of-the-art CMOS FDSOI technology using a dedicated FEOL and a simplified BEOL in 28FDSOI.

References

[1] Maurand R. et al., Nature Commun. 7, 13575 (2016).
[2] Vinet, M. et al., IEDM Tech. Dig. (2018).
[3] Petit, L. et al., Nature 580, 355–359 (2020).
[4] Pool C.P. et al., 2nd edition (1983).
[5] Koppens, F.H.L. et al., Nature 442, 766-771 (2006).
[6] Dehollain J. P. et al., Nanotechnology (2013).
[7] Jones, C. et al., Phys. Rev. X 8, 021058 (2018).
[8] Niquet Y.-M. et al., IEDM Tech. Dig. (2020).
[9] Hutin, L. et al., Device Research Conference (2021).
[10] Zwerver A.M.J. et al. arXiv:2101.12650 (2021).
[11] Bédécarrats, T. et al., IEDM Tech. Dig. (2021).

A Physics-based TCAD Framework for NBTI

Ravi Tiwari[1], Meng Duan[2], Mohit Bajaj[3], Denis Dolgos[4], Lee Smith[5], Hiu Yung Wong[6], and Souvik Mahapatra[1]
[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India
[2]Synopsys Northern Europe Ltd, Glasgow, UK, [3]Synopsys India Private Limited, Bangalore, India, [4]Synopsys Switzerland LLC, Zurich, Switzerland, [5]Synopsys Inc., Mountain View, CA, USA, [6]San Jose State University, San Jose, CA 95192, USA
*Phone: +91-222-572-0408, Email: souvik@ee.iitb.ac.in

Abstract: A physics-based framework is incorporated in TCAD to model the primary mechanisms responsible for Negative Bias Temperature Instability (NBTI) in P channel High-K Metal Gate (HKMG) MOSFETs. Three underlying mechanisms are treated including interface trap generation-passivation via a Reaction-Diffusion (RD) model and its charge occupancy via an Activated Barrier Double Well Thermionic (ABDWT) model, hole trapping and de-trapping in pre-existing defects in the gate stack are modeled via an ABDWT model, and bulk trap generation-passivation is modeled via a Reaction-Diffusion-Drift (RDD) model. The framework is used to model measured NBTI time kinetics for DC stress-recovery and various mixed DC-AC gate pulse segments for planar devices. Furthermore, the same framework is also used to test NBTI behavior in 3D FinFETs.

Introduction: The threshold voltage shift (ΔV_T) due to Negative Bias Temperature Instability (BTI) is an important issue in P channel HKMG MOSFETs [1]. The NBTI mechanism is controversial and various models are proposed [2]-[4]. The threshold voltage degradation (ΔV_T) is due to primary NBTI mechanisms, i.e., interface trap generation-passivation (giving rise to an interface trap charge density ΔN_{IT}) at or near the channel/interlayer (IL) interface and its charge occupancy (ΔV_{IT}), trap generation and passivation in the interlayer (giving rise to ΔV_{OT}), and charge trapping in pre-existing defects in the interlayer (giving rise to ΔV_{HT}) [6]. Previous NBTI modeling work in the 1D BTI Analysis Tool (BAT) is successfully applied to 2D/3D TCAD with the focus on trap generation and passivation (ΔN_{IT}) using an RD model [10][12] (using Synopsys tools [13][14]). In this work, we extend our NBTI framework in TCAD to improve the modeling of interface trap occupancy (ΔV_{IT}), charge trapping in the interlayer (ΔV_{HT}), and trap generation and passivation and its occupancy in the interlayer (ΔV_{OT}) (Fig.1). The advantages of modeling NBTI in TCAD are capturing quantum effects, source/drain induced mechanical strain impact in the channel which are essential for scaled devices [10]. The framework (Fig.1) is validated by modeling the measured data for pMOSFETs in 2D devices (Fig.4-10). Furthermore, the framework is used to examine the reliability behavior briefly in 3D FinFETs (Fig.11).

NBTI Models: The inversion layer hole tunnels to Hydrogen (H) passivated defects at the channel/interlayer interface during NBTI stress, reacts with H passivated defects, breaks, generates interface states, and releases H (Fig. 2a). The released H diffuses and reacts with another H passivated defect inside the interlayer to form H_2, which diffuses away. This is called the Reaction-Diffusion (RD) model. An anode hole injection (AHI) -related process triggered by a hole generates the trap in the interlayer and releases H, which further diffuses, reacts with another H passivated defect inside the interlayer, and generates defect states and H-species (H-molecule, H-ion). The generated H-molecule and H-ion diffuses towards the gate. This is called the Reaction-Diffusion-Drift (RDD) model for the traps generated in the interlayer [9] (Fig.2b). The ABDWT model is used for charge trapping in pre-existing defects in the interlayer (Fig.3).

TCAD framework: The device structure is generated in 2D (Fig.4a) and 3D (Fig.11a). In the device simulation [14], the defect dissociation by holes during BTI stress utilizes the Capture-Emission De-passivation (CED) model. A multi-state configuration (MSC) is defined for H-passivated defects at channel/interlayer and interlayer/HfO2 interfaces. The MSC-hydrogen transport degradation model is accounted for the reaction between mobile hydrogen species and localized states (H-passivated defects) together with the hydrogen transport. This is the RD model for interface traps in TCAD (Fig.2a). The RD model is combined with the ABDWT model for charge occupancy of generated traps (ΔV_{IT}), which previously has been done empirically via a Transient Trap Occupancy Model (TTOM) [10]. The RDD model utilizes a similar framework as the RD model (Fig.2b). Newly incorporated Hydrogen species (HydrogenSpeciesA/B/C, HydrogenIon) are utilized to isolate the AHI-related process triggered by holes for bulk trap generation in the interlayer (Fig.4b-d), which leads to the formation and diffusion of HydrogenSpeciesA/H_2/H_2^+ (Fig.2b). For charge trapping in interlayer bulk, the ABDWT model considers a trap with two states E_1 (uncharged) and E_2 (charged) connected through a thermally activated energy barrier E_B which is lowered by the applied bias (Fig.3). In the real scenario, traps are located inside the interlayer where carriers can be captured and emitted additionally through tunneling processes. The ABDWT model is effective, connecting both the processes in an effective manner by fit factors and distributions for E_B and E_2 [14]. The backend of 1µm is used to allow H_2 diffusion (Fig.4a).

Device and measurement details: Gate First (GF) N/P HKMG MOSFETs having ultra-thin thermal IL and HfO_2 HK are used [1]. IL scaling is done using thermal process tweak for D1 (3Å) with lower N%, and N based IL, D2 (1.5Å), EOT of HK is 4.6Å. DCIV data (Fig.5) are with delay correction [11]. The ΔV_T (Fig.7-10) is measured with a 10µs delay [5], [6].

TCAD Modeling: The modeling of DCIV measured ΔN_{IT} kinetics at different gate bias (V_G) and temperatures (T) are shown for the D1 device using the RD model (Fig.5). The measured ΔV_T time kinetics is modeled with its subcomponents at fixed V_G/T during stress (Fig.6a) and recovery (Fig.6b). The overall ΔV_T time kinetics modeling is shown for mixed V_G/T during stress (Fig.7a) and bias dependence of fixed time ΔV_T at the end of 1Ks stress at different temperatures (Fig.7b). The modeling of ΔV_T recovery at various measurement conditions, i.e., the recovery bias ($V_{G\text{-}REC}$) of 0V (Fig.8a), the $V_{G\text{-}REC}$ dependence after longer (1Ks) stress time (Fig.8b), the stress time-dependent recovery behavior (Fig.8c), and the $V_{G\text{-}REC}$ dependence after shorter stress time of 100ms (Fig.8d) is also achieved. It is appreciable that TCAD can mimic the re-stressing during recovery after a short stress time (100ms). The model framework shows good agreement with experimental data for the D2 device (Fig.9a-d) showing DC stress-recovery bias variation, DC-AC-DC, AC-DC-AC, and pure AC gate pulses having variable pulse duty cycle (PDC). The framework is extended to 3D FinFETs (Fig.10a). The simulated trap profile and charge trapping are shown (Fig.10b-e). Time kinetics of the generated traps and trapped charges in the FinFET show similar behavior as in 2D TCAD simulations for planar devices, indicating the 3D capability of the framework (Fig.10f-g).

Conclusion: A fully physical TCAD framework to model NBTI is validated with measured data of differently processed MOSFETs by incorporating the RD model with the ABDWT model (as TTOM), the RDD model, and the ABDWT model for the interface traps, bulk traps, and charge trapping, respectively. TCAD framework validation in 3D with measured data for a FinFET and GAA-FET is in progress.

References: [1] Joshi IRPS 2013, [2] Mahapatra, TED 2013, [3] Rzepa, MR 2018, [4] Stathis, MR 2018, [5] Mukhopadhyay, TED 2017, [6] Parihar, TED Feb 2018, [7] Mahapatra, TDMR 2020, [8] Nilotpal IRPS 2020, [9] Tarun, TED, Feb 2021, [10] Tiwari, TED May 2019, [11] Mukhopadhyay, IRPS 2014, [12] Tiwari, SISPAD 2020, [13] Sentaurus Process T-2022.03, [14] Sentaurus Device T-2022.03.

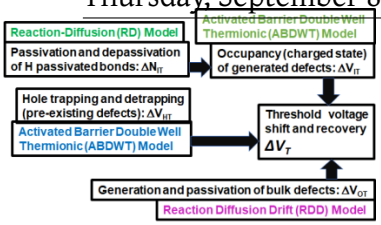


Fig. 1. Schematic of TCAD framework for NBTI during stress and recovery for High-K Metal-Gate devices. The ΔV_T subcomponents and corresponding models are shown.

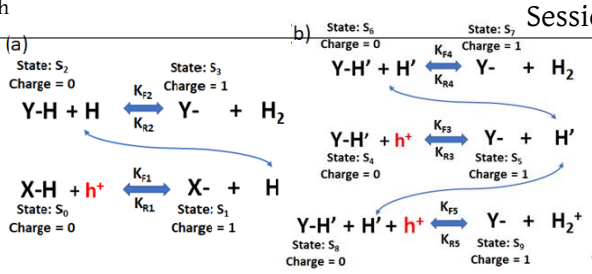


Fig. 2. Multi-State Configuration (MSC)-hydrogen transport degradation model, utilizing Capture Emission De-passivation model for defect dissociation used in (a) the RD model for interface traps generation-passivation, and (b) the RDD model for bulk trap generation-passivation in TCAD. H': HydrogenSpeciesA, H_2^+: Hydrogen Ion, H: Hydrogen Atom, H_2: Hydrogen Molecule, h^+: hole.

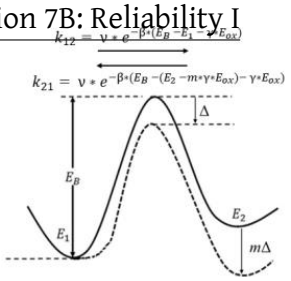


Fig. 3. Schematic of the ABDWT model in TCAD to capture hole trapping de-trapping in pre-existing defects, showing thermally activated barrier E_B, with forward and backward transition rates [8].

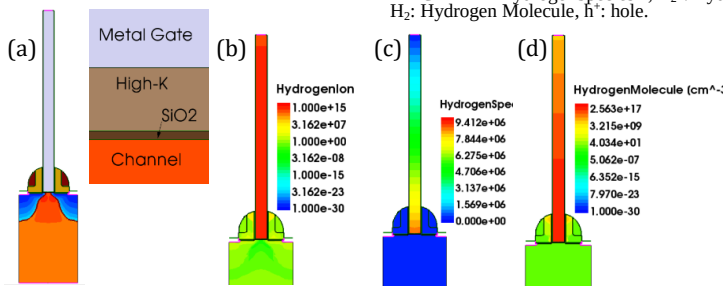


Fig. 4. Device showing gate stack for (a) 2D pMOSFET, and diffusion of different hydrogen species for (b) (c) (d) 2D MOSFET used in RDM and RDDM after 1Ks stress time.

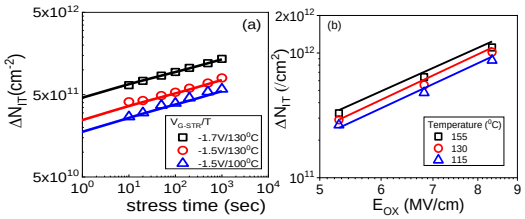


Fig. 5. TCAD modeling of DCIV measured (a) time kinetics of ΔN_{IT}, (b) field dependence of ΔN_{IT} at different temperatures after fixed stress time of 1Ks for the D1 device. Line: TCAD simulation, symbol: measured data.

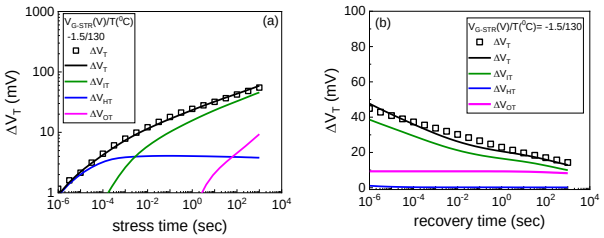


Fig. 6. TCAD modeling of UF-MSM measured ΔV_T of the D1 device during (a) stress, and (b) recovery time kinetics. ΔV_T subcomponents are also shown. Line: simulation, symbol: measured data.

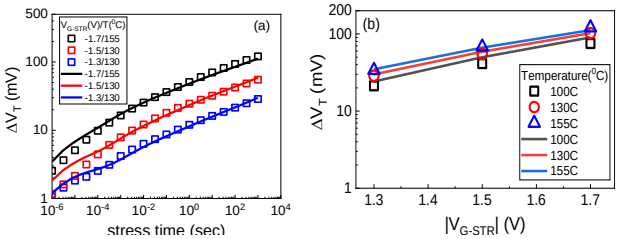


Fig. 7. TCAD modeling of UF-MSM measured ΔV_T of the D1 during (a) stress time kinetics at mixed stress voltage/temperature ($V_{G\text{-}STR}$/T), (b) $V_{G\text{-}STR}$ dependence of fixed time ΔV_T at different T. Line: simulation, symbol: measured

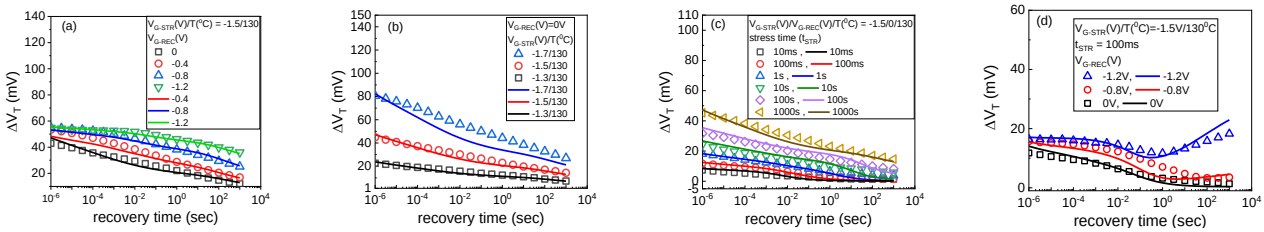


Fig. 8. TCAD modeling of UF-MSM measured ΔV_T of the D1 device showing (a) recovery at 0V reccovery bias, (b) recovery bias dependence, (c) stress-time dependence, and (d) recovery bias dependence after short stress time of 100ms. Line: TCAD simulation, symbol: measured data.

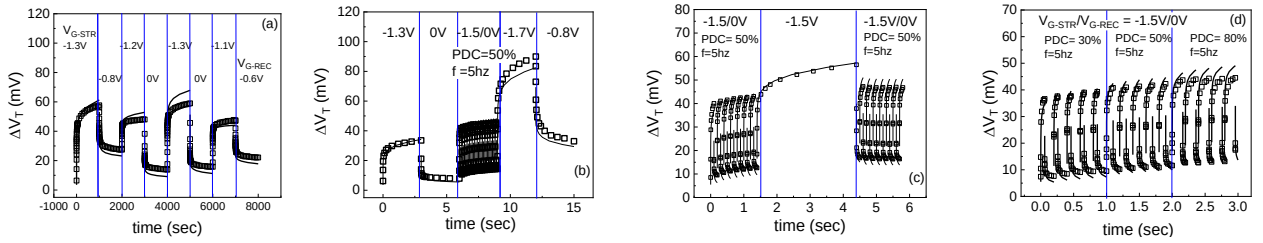


Fig. 9. TCAD modeling of measured ΔV_T of the D2 device (a) DC segments with varying $V_{G\text{-}STR}/V_{G\text{-}REC}$, (b) (c) mixed DC-low frequncy (f) AC, and (d) low f AC cycle with variable PDC. Black Line: TCAD simulation, symbol: measured data.

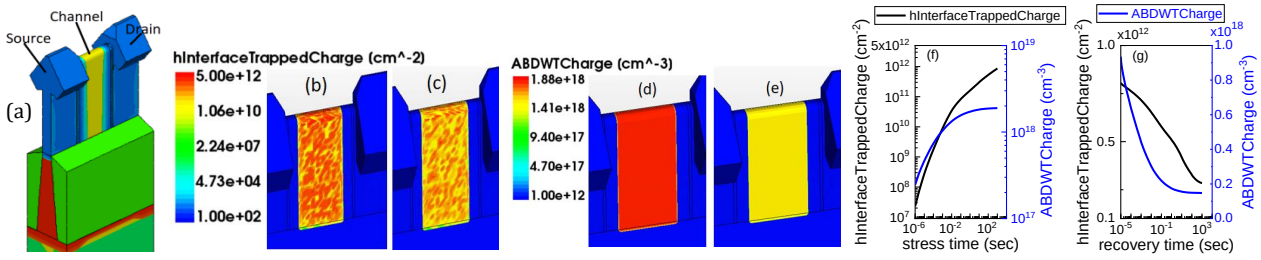


Fig. 10. 3D isometric view of FinFET showing (a) device structure, generated traps in the fin area using the RD and the RDD model at the end of (b)1Ks stress, (c) 1Ks recovery, charge trapping in the fin area using ABDWT model at the end of (d)1Ks stress, (e) 1Ks recovery. Corresponding time kinetics is shown during (f) stress, and (g) recovery.

A Stochastic Simulation Framework for TDDB in MOS Gate Insulator Stacks

Satyam Kumar[#], Tarun Samadder[#], Dimple Kochar and Souvik Mahapatra[*] ([#]Equal contributions)
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India
[*]Phone: +91-222-572-0408, Email: souvik@ee.iitb.ac.in

Abstract: **A stochastic simulation framework is proposed for Time Dependent Dielectric Breakdown (TDDB), comprising of Reaction-Diffusion-Drift (RDD) model for generation of bulk traps and cell-based percolation model for initiating breakdown. Stochastic, and computationally lesser expensive deterministic implementations of RDD model are shown as identical. Trap time kinetics calculated from the latter, with variation in one parameter for stochasticity, is used in the percolation model. The framework provides a physical justification for trap time kinetics change as oxide thickness (T_{OX}) is varied. Model is validated using measured data from published sources. The framework can consistently explain T_{OX} dependence of Weibull distribution slope (β) and time to breakdown (t_{BD}) mean at different stress gate bias (V_G) and temperature (T).**

Introduction: TDDB is attributed to formation of a percolation path by generated gate insulator traps (ΔN_{OT}) during gate stress at chosen V_G / T, and resulting sudden increase in gate leakage [1]. The underlying physical processes for TDDB are stochastic in nature, t_{BD} shows a Weibull distribution with slope β, which reduces non-linearly as T_{OX} is scaled (Fig.12, compilation from [1]-[4]). Voltage Acceleration Factor (VAF), or the slope of V_G dependence of mean t_{BD}, reduces at higher T (Fig.13) [5], [6] but increases as T_{OX} is reduced (Figs. 14, 15) [6], [7]. The existing modelling efforts are addressed towards percolation framework (using analytically computed trap kinetics) [1], [8], and impact of T and T_{OX} on VAF [9]-[11]. The non-linearity of Weibull β versus T_{OX} is not explained.

TDDB simulator: This work proposes a simulation framework, with RDD model for trap generation and cell-based percolation model for breakdown, Fig.1, and can explain key experimental features (Figs. 12-15). In RDD model, dissociation of Hydrogen (H) passivated defects is triggered at a stress V_G / T, released H atoms diffuse and react to dissociate other defects and release of molecular H_2 and ionic (H_2^+, OH^-) species; diffusion / drift of these species governs long-time kinetics of ΔN_{OT} (~ dissociated defects), Fig.2. Forward / reverse reaction rates for atom (K_{F1} / K_{R1}), molecule (K_{F2} / K_{R2}) and ion (K_{F3} / K_{R3}), diffusivities of H and H_2 and drift of H_2^+, OH^- are Arrhenius T activated, K_{F1} and K_{F3} are the only adjustable parameters and determine magnitude and long-time slope (*n*) respectively of the ΔN_{OT} time kinetics. RDD model is implemented by Gillespie Stochastic Simulation Algorithm (GSSA) [12], [13], Fig.3. A deterministic version by solving partial differential equations [14] is also used. The cell-based percolation framework, Fig.4, is same as [8], however, it uses ΔN_{OT} kinetics as per RDD model, with T_{OX} dependent *n*.

Stochastic RDD model: 3-D simulation domain is divided into cells, and consists of gate insulator and backend. Precursors are randomly assigned to channel/gate insulator interface cells (1st interface, generation of atoms) and in cells in one plane inside gate insulator (2nd interface, generation of molecules and ions). Reaction (forward/reverse for 3 cases), 3-D diffusion / drift for 3 species and lock-in / un-lock in (for molecule) give a total of 26 events. GSSA generates 2 random numbers, to select one of the 26 events (based on their individual propensities), and time stamp for next event (based on the inverse of total propensity), Fig.3. The selected event is executed, propensities and time are updated, and loop is iterated till stipulated time. Areal density of generated traps at both interfaces is converted into a volume density of traps and feed into percolation model - an equivalent deterministic model is validated with stochastic results (mean) and is used to reduce execution time (since stochastic model is computationally expensive, especially for long stress time). K_{F1} is related to stress V_G / T, random variation in K_{F1} (around a mean) for the deterministic model reproduces stochasticity, and reduction in K_{F1} after stress initiates recovery. Measured mean t_{BD} at various V_G / T across T_{OX} is used to validate RDD model.

Percolation model: 3-D simulation domain (gate insulator) is divided into cells. Traps generated at a constant rate (~ slope *n*) is randomly placed in these cells, time is noted, and formation of percolation path (from channel to gate) is checked; procedure is repeated in time till a path is formed (multiple configurations are considered for breakdown), Fig.4. t_{BD} area scaling (see [1]) is used to validate the percolation framework.

Results: The ion to molecule ratio (adjusted by K_{F3}) determines *n* during stress and fraction remaining (FR) during recovery of simulated ΔN_{OT} time kinetics. Molecule (Fig.5) and ion (Fig.6) dominated (after 2nd interface reaction) systems show *n* / FR of ~ 0.17 / 0.8 and ~ 0.5 / 1 respectively, intermediate *n* (~ 0.3) is obtained with mixed molecule-ion system (Fig.7), their ratio is varied to obtain varying *n* / FR between two extremes (Fig.8). Figs. 5-7 show individual stochastic simulations, their mean and deterministic simulation, similarity of the two implementations is shown in Fig.8. Location of particles from a single stochastic simulation shows symmetrical (asymmetrical) vertical distance covered for molecules (ions) during recovery (Fig.9). Drift of ions during stress but diffusion after stress (in absence of electric field) result in this asymmetry, and lack of recovery after stress for ion dominated system (Fig.6).

Validated deterministic RDD model is used to generate ΔN_{OT} time kinetics at different V_G, showing power law dependence at long-time (Fig.10). Two different VAF definitions are used, one (VAF-TDDB) to attain a particular ΔN_{OT} related to critical trap density for breakdown (N_{BD}), other (VAF-FT) governs spacing between curves at fixed time; at long time (relevant for TDDB), both are related by *n* (VAF-TDDB = VAF-FT / *n*). Percolation simulation results in Weibull distribution for t_{BD} (Fig.11), with mean t_{BD} dependence on V_G / T (RDD model) and β related to *n* ($\beta = n * T_{OX} / a_0$, with a_0 being cell dimension, Fig.4, as in [8]).

Non-linear reduction of β with T_{OX} scaling is modeled using monotonic reduction in both *n* and a_0 (Fig.12). Variation of *n* in the range of ~ 0.5 to 0.25 for T_{OX} scaling (~ 11nm to 1nm) can be reproduced by ion dominated to mixed molecule-ion RDD simulations. Presence of larger ion proportion at thicker T_{OX} is presumably due to higher stress V_G (~ higher energy of injected electrons and holes in the gate insulator) during TDDB stress.

RDD model is quantified by modeling measured mean t_{BD} as a function of V_G at different T for 2 different T_{OX} (Fig.13), and t_{BD} versus V_G at different T_{OX} (Fig.14). K_{F3} is varied to obtain *n* for a particular T_{OX} (as per Fig.12). K_{F1} is calculated by Eq. (1), Fig.2; polarization factor α (fixed value for all T_{OX}) necessitates the modeling of T dependence of VAF [9]. A fixed VAF-FT is able to reproduce increase in VAF-TDDB at lower T_{OX} (Fig.15) due to reduction in *n* (consistent with β versus T_{OX} variation).

Conclusions: RDD model provides a physical justification for *n* reduction at lower T_{OX}, to explain key experimental TDDB features as reduction in Weibull β and increase in mean TDDB-VAF at lower T_{OX}. Equivalence of stochastic and deterministic RDD models imply the latter is useful for long-time simulation, for mean t_{BD} calculation at different V_G and T.

References: [1] Degraeve, TED,1998. [2] Wu, TED, 2002. [3] Stathis, IEDM, 1998. [4] Nicollian, IEDM, 2006. [5] Wu, JAP, 2013. [6] Wu, TED, 2009. [7] Wu, IRPS, 2009. [8] Nigam, IRPS, 2009. [9] McPherson, IEDM, 1998. [10] DiMaria, JAP, 1995. [11] Alam, TED, 2002 [12] Gillespie, J. Phys. Chem, 1977. [13] Kumar, TED, 2020. [14] Samadder, TED, 2021.

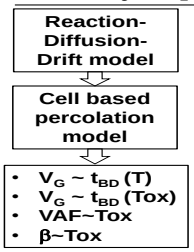


Fig. 1: Schematic of TDDB simulator- Percolation model produces t_{BD} distribution while RDD model calculates bulk trap kinetics to explain various dependencies as listed at bottom.

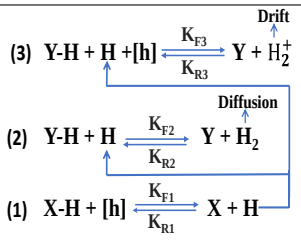


Fig. 2: Schematic of RDD model chemical equations showing hole assisted release of atomic hydrogen, it's dimerization into molecular hydrogen and generation of ionic species.

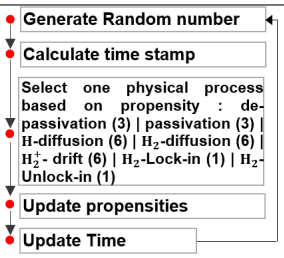


Fig. 3: Stochastic RDD model implementation using Gillespie stochastic simulation algorithm (SSA). Individual propensities related to a process are indicated in brackets.

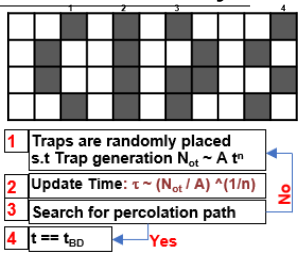


Fig. 4: Schematic of KMC percolation model showing trap generation at a constant rate n and breakdown time t_{BD} calculation once percolation path is formed.

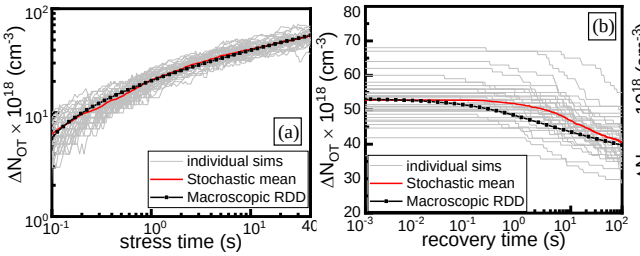


Fig. 5: Stochastic individual and mean time kinetics, matched with deterministic RDD framework for only molecule case during (a) stress, time exponent ~ 0.17 (b) recovery, fraction remaining ~ 80%.

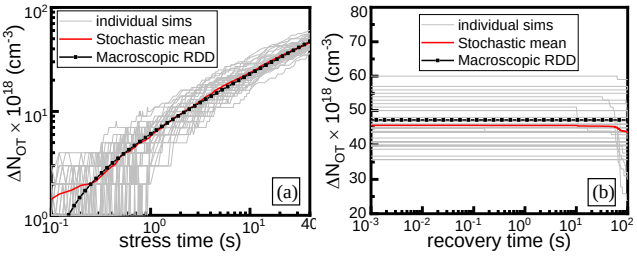


Fig. 6: Stochastic individual and mean time kinetics, matched with deterministic RDD framework for only ion case during (a) stress, time exponent ~ 0.5 (b) recovery, fractional remaining ~ 100%.

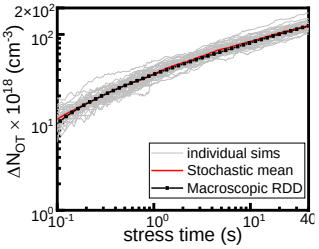


Fig. 7: Stochastic individual and mean stress kinetics having time exponent of ~ 0.3, matched with deterministic RDD model for mixed molecule - ion case.

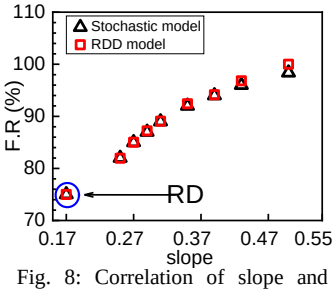


Fig. 8: Correlation of slope and recovery for different ion to molecule ratio from stochastic mean and deterministic simulation. Reproduced from [11]

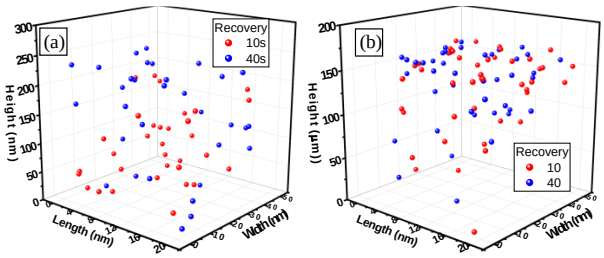


Fig. 9: Location of particles in the backend of structure during recovery for two different time stamps (a) symmetrical location of molecular species along the vertical direction (b) asymmetrical location of ionic species along the vertical direction.

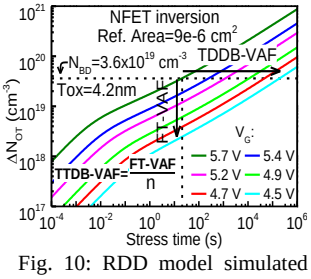


Fig. 10: RDD model simulated bulk trap (ΔN_{OT}) kinetics at different stress voltages (V_G) for 4.2nm oxide. Data from [6] is area scaled to 9x10^{-6} cm^2.

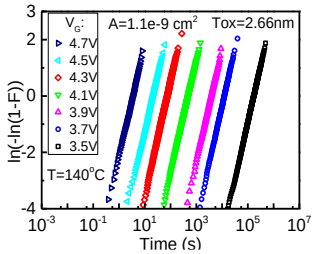


Fig. 11: Percolation model simulated Weibull CDF for different stress voltages (V_G) for 2.66nm oxide. Data from [2].

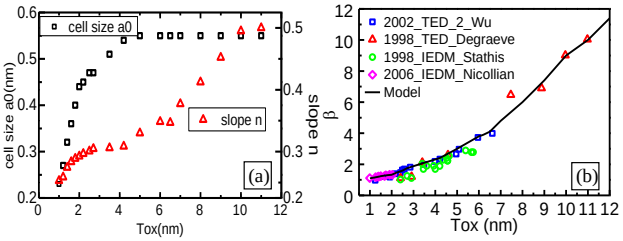


Fig. 12: (a) Oxide thickness (Tox) dependence of cell size (a0) and bulk trap slope (n). (b) Oxide thickness (Tox) dependence of Weibull slope (β) from various literatures (see legend)

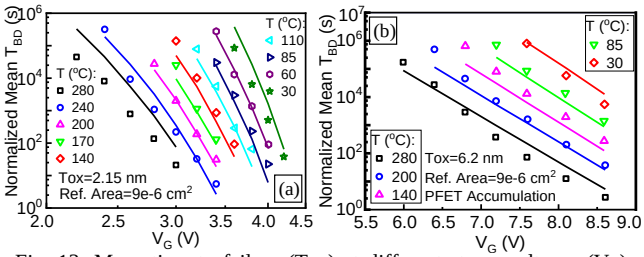


Fig. 13: Mean time to failure (T_{BD}) at different stress voltages (V_G) and temperatures for (a) a thin (2.15nm) oxide under substrate injection mode of a PMOS capacitor. Data from [5]. (b) for a thick oxide (Tox=6.2nm) in PFET accumulation. Data from [6].

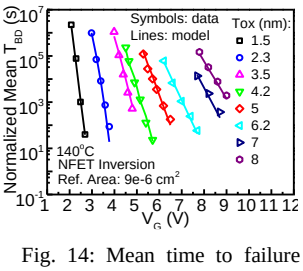


Fig. 14: Mean time to failure (T_{BD}) versus stress voltage (V_G) for various oxide thickness (Tox). Data from [6], [7].

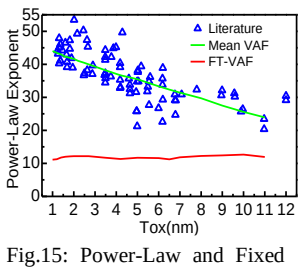


Fig.15: Power-Law and Fixed Time (FT) VAF dependence on oxide thickness (Tox) from literatures. Data from [7].

CARAT - A Reliability Analysis Framework for BTI-HCD Aging in Circuits

Prasad Gholve[1*], Payel Chatterjee[1*], Chaitanya Pasupuleti[1#], Hussam Amrouch[2#], Narendra Gangwar[1#], Shouvik Das[1#], Uma Sharma[1#], Victor M van Santen[2#], and Souvik Mahapatra[1] (*equal contributions, #listed alphabetically)

[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India
[2]Institute of Computer Architecture and Computer Engineering, University of Stuttgart, Germany
*Phone: +91-222-572-0408, Email: souvik@ee.iitb.ac.in

Abstract: Circuit Aging Reliability Analysis Tool (CARAT), a framework that calculates random activity (frequency and duty) aware degradation of FETs to simulate circuit aging under real operating workloads is proposed. Bias Temperature Instability (BTI) and Hot Carrier Degradation (HCD) induced degradation of FETs is calculated in a cycle-by-cycle manner based on actual terminal waveforms grabbed from SPICE. Framework capability is demonstrated by using Level Shifter (LS) under random data-path activity, and Ring Oscillator (RO) under Dynamic Voltage Frequency Scaling (DVFS) conditions. The risk associated with standard blanket approach is discussed.

Introduction: NBTI and HCD remain as key reliability concerns in FETs [1], [2], thereby affecting various circuits [3]. Circuit aging can be estimated by assigning a blanket number to all FETs (DC, or AC with some pre-defined duty). Alternatively, activity awareness can be found using an Age-based approach [4], which relies on finding an effective duty and suitable compact models for FET degradation. Finding effective duty can be challenging under realistic data-path workloads, more so under DVFS, while compact modelling is a challenge for BTI due to recovery, which is complex in nature. The presence of Self Heating (SH) further complicates the effective duty approach [5], since it depends on the actual number of on/off transitions.

CARAT uses cycle-by-cycle simulation to determine BTI and HCD in FETs. BTI is calculated by a physics-based BTI Analysis Tool (BAT) [6]. Hot-carrier Empirical Analysis Tool (HEAT), a compact model including time transformation concept is used for HCD (since it has no recovery) [7]. In our earlier work, CARAT is used to analyse large SRAM array under actual workloads [8]. In this work, the working of CARAT is described, with examples to demonstrate the need for cycle-by-cycle analysis.

Framework description: CARAT presently is a standalone tool that invokes SPICE, Fig.1. It requires circuit Netlist and Model Cards for FETs. For this work, we used HSPICE [9] and BSIM-CMG model [10] for FinFETs (calibrated with device data [11]). It has a fully automated Control Framework that runs HSPICE, grab terminal waveforms and temperature (T) due to SH for each FET (Grabber), shape waveforms (Pulse Shaper) suitable for BTI and HCD analysis, Fig.2 (a), runs BAT and HEAT for short time (*e.g.*, 1µs), extrapolate (Extrapolator) short-time individual FET degradation to End-of-Life (EOL), update (MC Updater) Model Cards of each FET, runs HSPICE, and compare initial and final runs. It handles process variability, BTI and HCD variability, by having individual BSIM, BAT and HEAT ModelCards for each FET. The extrapolation can be done at one go or in several loops (with intermediate CARAT runs) to EOL. Only V_T-shifted aging is used at present (V_T: threshold voltage), although other model parameters can also be aged in the MC Updater module (by using suitable correlation between different parameters [11]).

Grabbed cycle-by-cycle waveform is shaped as trapezoidal on /off phases for BAT, and staircase transition phases for HEAT, Fig.2, a low frequency inverter simulation is used in this case for demonstration. Parameters for BAT and HEAT are obtained by calibration against device data [6], [7]). BTI is considered for PFET (it is negligible for NFET [12]), and HCD for both FETs. BTI shows recovery but HCD does not. High frequency inverter simulation for short time is shown, with resulting BTI at different duty, Fig.3, and HCD at different frequency, Fig.4. BTI increases at higher duty but is frequency independent [1], HCD increases with larger number of on/off transitions due to higher frequency. Extrapolation is done by using iso-bias DC simulation reference, after shifting it vertically (Y-axis) over BTI and laterally (X-axis) over HCD short-time AC kinetics. A simple time-power law is not accurate for projection to EOL (as is usually done [13]), since the actual kinetics is substantially different. FET ModelCards are updated, HSPICE is run again, and pre- and post- aging runs are compared, Fig.5. In this case, rising edge is more degraded than falling edge, since PFET has both BTI and HCD and NFET has only HCD. Standard inverter-based RO simulations show similar BTI across different stages (BTI is frequency independent), but higher HCD for lower number of stages (higher frequency and number of transitions), Fig.6, consistent with measurements [14].

Activity analysis: A LS circuit, Fig.7, is used. Table-I lists short time BTI and HCD degradation of all FETs, under different data-path like random activities, Fig.8. All cases have effective 50% duty for 1µs simulation time: (A) is simple stress and recovery, (B) and (C) are AC with different frequency but same duty (only a fraction of the entire pulse duration is plotted), (D) through (F) have mixed frequency-duty combinations. Due to difference in bias rail (V_{DD}), in general, FETs M1-M5 degrade lower than M6-M10. However, for any particular waveform, the BTI and HCD of different FETs can be similar or different, hence assignment of a single value (based on V_{DD}) is ineffective. In spite of same effective duty, for Case-A, BTI is highest (longest on/off phases) but HCD is lowest (lowest number of transitions), while Case-C HCD is highest (maximum number of transitions) and BTI is lowest (shortest on/off phases) among different cases studied. All other waveforms are in between these two, and BTI and HCD are respectively determined by sequence and duration of subsequent on/off phases and number of transitions. This makes the effective duty approach ineffective.

DVFS analysis: RO (7-stage) is subjected to different DVFS like V_{DD} waveforms, Fig.9, and resulting BTI and HCD degradation are obtained, and compared to standard analysis using minimum, average and maximum V_{DD} (of DVFS waveform) for the entire duration. Actual case degrades differently from standard analysis and cannot be reproduced by taking the mean (of maximum and minimum) V_{DD} simulations or from average V_{DD} simulations. It is challenging to reproduce DVFS like scenario with an effective V_{DD} approach.

Conclusion: Realistic data-path like input waveforms and DVFS pose severe challenges for circuit simulation. Such situations are not representable by an equivalent duty and V_{DD} simulations, and therefore, cycle-by-cycle simulations are required. The proposed CARAT framework achieves the same. Due to lack of recovery, a suitable compact model can handle HCD under arbitrary time / V_{DD} segments. Compact model development is very challenging for BTI due to the presence of recovery, which can become very complex under arbitrary time / V_{DD} segments. Hence, a physical framework, BAT, is used. The entire CARAT framework is fully automated and implemented in parallel mode that significantly reduce run time. At present it is a standalone tool, integrating it inside SPICE would reduce the file size limitations (for grabbed waveforms) and further improve the runtime.

References: [1] A. Rahman, IRPS, 2017 [2] Qu, EDTM, 2017,[3] Mahapatra, IRPS, 2014, [4] Mishra, TED, 2019,[5] Thirunavukkarasu, TED, 2019, [6] N. Parihar, TED 2018, [7] U. Sharma, TED, 2019 [8] Van Santen, IRPS 2020 [9] Synopsys HSPICE M-2017.03, [10] BSIM-CMG manual, 103.0, [11] S. Mishra, TED 2019, [12] S. Ramey, IRPS 2013, [13] B. Tudor, ICSICT 2010[14] A. Kerber, EDTM 2017

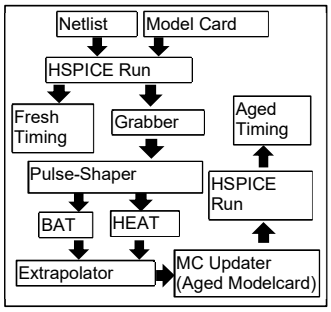


Fig.1. Schematic illustration of CARAT simulation flow (Control Framework).

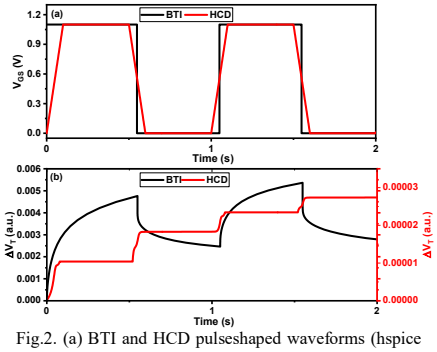


Fig.2. (a) BTI and HCD pulseshaped waveforms (hspice extracted coincides with HCD) (b) CARAT simulated $\Delta V_{T\text{-}BTI}$ and $\Delta V_{T\text{-}HCD}$ time kinetics for two cycles.

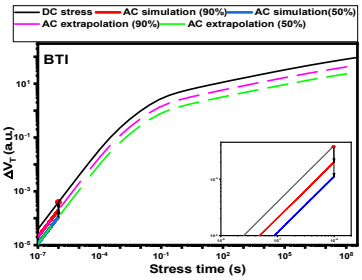


Fig.3. BTI extrapolation for 50% and 90% duty input waveform, with AC simulation till 1 µS and extrapolation to 10 years. (Inset: BTI Extrapolation step from DC to AC at 1uS plotted).

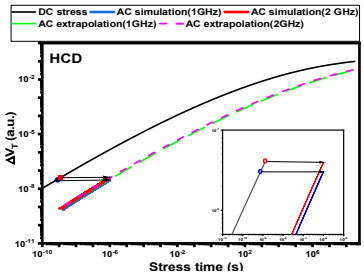


Fig.4. HCD extrapolation for 1GHz and 2GHz frequency input waveform, with AC simulation till 1 µS and extrapolation to 10 years. (Inset: HCD Extrapolation step from DC to AC at 1uS plotted).

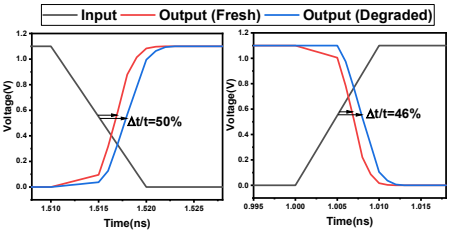


Fig.5. Input and output waveforms without and with aging, showing delay in (a) rise time and (b) fall time for Inverter.

Ring Oscillator	Frequency	BTI Degradation at 1µs (V)	HCI Degradation at 1µs (V)
7 Stage	16.76 GHz	1.13E-7	1.2E-7
11 Stage	10.66 GHz	1.16E-7	1.03E-7
21 Stage	5.58 GHz	1.15E-7	8.4E-8
31 Stage	3.78 GHz	1.15E-7	7.4E-8
51 Stage	2.29 GHz	1.15E-7	6.3E-8
101 Stage	1.16 GHz	1.15E-7	5E-8

Fig.6. BTI and HCD degradation for different Ring Oscillator stages at 1µS

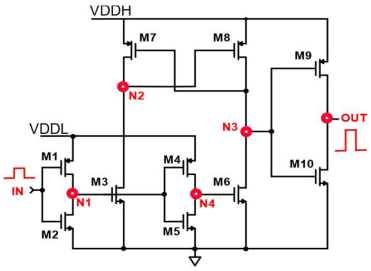


Fig.7. Schematic of Level Shifter Circuit used to demonstrate activity aware degradation of transistors with CARAT.

Table-1. Degradation of transistors obtained for input waveforms shown in Fig.6 indicating activity aware degradation of transistors in Level Shifter. All simulations done till 1µS with ΔV_{TH} values corresponding to BTI and HCD in Volts.

Transistor	BTI (A)	HCD (A)	BTI (B)	HCD (B)	BTI (C)	HCD (C)	BTI (D)	HCD (D)	BTI (E)	HCD (E)	BTI (F)	HCD (F)
M1	5E-09	1E-09	2.9E-09	2.6E-09	2.9E-09	4E-09	4.4E-09	2.7E-09	2.7E-09	2.7E-09	4.4E-09	2.4E-09
M2	0	1.6E-09	0	2.3E-09	0	3.1E-09	0	2.2E-09	0	2.2E-09	0	2E-09
M3	0	2.9E-09	0	1.2E-08	0	1.9E-08	0	1.2E-08	0	1.2E-08	0	1.1E-08
M4	4.8E-09	1.4E-09	2.9E-09	2.6E-09	2.9E-09	4.1E-09	3.5E-09	2.7E-09	2.9E-09	2.7E-09	3.5E-09	2.4E-09
M5	0	1.2E-09	0	2.5E-09	0	3.8E-09	0	2.6E-09	0	2.6E-09	0	2.3E-09
M6	0	2.8E-09	0	1.3E-08	0	2E-08	0	1.3E-08	0	1.3E-08	0	1.2E-08
M7	7.2E-07	8.6E-09	1.2E-07	3.8E-08	1.2E-7	6.2E-08	1.4E-7	4.1E-08	1.5E-7	4.1E-08	1.4E-7	3.6E-08
M8	2E-07	3.2E-09	1.2E-07	3.8E-08	1.1E-7	5.9E-08	1.2E-7	3.9E-08	1.2E-7	3.9E-08	3.1E-7	3.4E-08
M9	7.2E-07	2.5E-09	1.2E-07	9.2E-09	1.2E-7	1.4E-08	1.4E-7	9.5E-09	1.5E-7	9.5E-09	1.4E-7	8.3E-9
M10	0	2.2E-09	0	8.3E-09	0	1.2E-08	0	8E-09	0	8E-09	0	6.9E-9

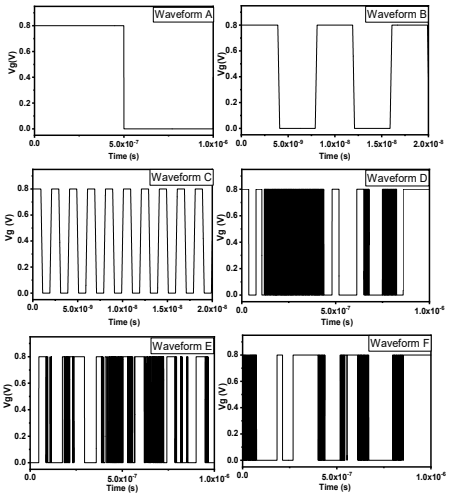
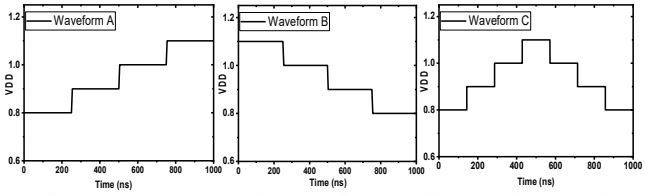


Fig.8. Input waveforms used to demonstrate activity aware degradation of transistors with CARAT. (Waveform B and C although only shown upto 20nS in the figure are repetitive and applied for 1µS.)



Waveform	Degradation	CARAT	Minimum VDD	Average VDD	Maximum VDD
Waveform A	BTI	4.02E-08	2.9E-09	1.9E-08	1.1E-7
	HCD	9.2E-8	2.3E-8	5.2E-08	1.2E-7
Waveform B	BTI	3.81E-08	2.9E-09	1.9E-08	1.1E-7
	HCD	6.4E-8	2.3E-8	5.2E-08	1.2E-7
Waveform C	BTI	2.82E-08	2.9E-09	1.4E-08	1.1E-7
	HCD	7.3E-8	2.3E-8	4.6E-8	1.2E-7

Fig.9. Schematic of DVFS waveforms (top) and resultant degradation in (V) of each FET in 7-stage RO with CARAT, min, max and average analysis.

Trap and Self-Heating Effect Based Reliability Analysis to Reveal Early Aging Effect in Nanosheet FET

Sunil Rathore, Rajeewa Kumar Jaisawal, P. N. Kondekar, and Navjeet Bagga*
VLSI Design and Nano-scale Computational Lab, Electronics and Communication Engineering Department, PDPM-IIITDM Jabalpur, India- 482005 (*Email: navjeet@iiitdmj.ac.in)

Abstract: The reliability of the CMOS devices is severely affected due to the presence of interface (*Si-SiO₂*) trap charges and self-heating effect (SHE). In this paper, we investigated the trap and temperature-dependent performance barrier and aging issues in Nanosheet FET (NSFET). Through extensive TCAD simulations, we found that: a) the type (donor/acceptor) and concentration of the trap charge significantly modulate the device's threshold voltage (V_{th}); b) the location of the trap charges around the conduction band (CB) and valence band (VB) plays a crucial role; c) the performance degradation is observed in NSFET due to the ambient temperature (T_A) and SHE; d) the trap assisted SHE significantly affects the performance metrics viz I_{ON}, I_{OFF}, subthreshold slope (SS); e) the overall effect dominantly alters the V_{th} shift of the device, resulting in device aging. Hence, the analysis of trap assisted SHE by varying the ambient temperature is essential for reliable NSFET operation.

Introduction: To achieve higher device density, the downscaling of the device dimensions leads to the evolution of the semiconductor industry from planar devices to non-planar devices like gate-all-around FET, FinFETs, etc. [1]-[2]. Among all the promising candidates, the Nanosheet FET has gained significant popularity as a potential high-performance sub-5nm node device due to the higher drive current, improved electrostatic integrity, immune to short channel effects, and FinFET compatible layout [3]-[4]. However, the confined geometry of the NSFET is severely influenced by the performance barrier parameters like trap charges, self-heating effect, temperature, etc., which draw attention to research. The interface trap charges and induced SHE primarily affects the V_{th} of the device, which is an essential device design parameter. The shift in V_{th} can be used to decide the aging of the device, i.e., end of life (EOL) when V_{th} is shifted by ~50mV. *The key contribution of work is*: (*i*) the impact of interface trap charge in the *Si-SiO₂* interface is investigated by considering both donor and acceptor traps separately; (*ii*) the significance of the location where the trap charges lie in the energy bandgap (E_g) or around CB/VB is studied; (*iii*) the investigation of the influence of ambient temperature and SHE induced performance degradation with the presence of traps; (*iv*) evaluation of the aging effect (ΔV_{th}=50mV) for deciding the optimal reliability of the device.

Device Structure and TCAD Setup: The vertically stacked three-sheet NSFET (Fig.2a-b) is considered a baseline reference for the trap and SHE analysis using *Sentaurus* TCAD [5]. The source/drain (S/D) pads and channel regions are uniformly doped. The Gaussian doping is used in the extension region for realistic operation. The device parameters used in the simulation are mentioned in Table-I. Fig.2c shows the transfer characteristics (I_{DS}-V_{GS}), showing a good match in TCAD and experimental data [6]. The conventional drift/diffusion (DD) model is used to capture the carrier transport in the simulation setup. The density gradient model was invoked to capture the spatial and electrostatic quantum confinement in the 5nm thick channel region. The modified local density approximation model (MLDA) is incorporated to consider the carrier distribution near the oxide/silicon interface. The IAL mobility model governs the surface roughness scattering effect on mobility. The Slotboom and SRH models are used to capture the doping/temperature-dependent bandgap and recombination, respectively. Further, the coupled hydrodynamic and thermodynamic models are used to capture SHE-induced thermal degradation.

Results and Discussion: This work investigates the influence of interface trap charges and SHE-induced thermal effect on NSFET aging. In NSFET, the channel regions are surrounded by the low thermal conductivity material, i.e., *SiO₂*, which confines the heat flux in the channel direction (Fig.3a). This causes an increase in the lattice temperature, resulting in SHE-induced performance degradation (Fig.3b). To investigate the impact of interface trap charges, we considered both donor and acceptor traps at the *Si-SiO₂* of each channel in NSFET (Fig. 1). The location where the trap charges lie, i.e., either in between the E_g, above CB, or below VB, plays a vital role (Fig.4). The shift in location is fixed by the Fermi potential, i.e., $\varphi_F = kT/q\ ln(N_{ch}|n_i)$. The acceptor traps become negatively charged below the Fermi Level (FL) and neutral above the FL. Thus, the acceptor traps near the CB edge reduce the V_{th} (Fig.4a-c). In contrast, the donor traps become positively charged above the FL and neutral below the FL. Therefore, an opposite trend in V_{th} is observed with donor traps (Fig. 4d-f). These position-located trap charges shift the V_{th} and alter other parameters like I_{ON} and SS. The peak energy density of the traps follows the Gaussian distribution with a standard deviation (σ) for the donor and acceptor traps. By keeping the fixed trap concentration and σ, the impact of SHE by varying the ambient temperature has been analyzed (Fig.5). The increase in temperature shows a significant shift in V_{th} compared to the baseline (BL) case, i.e., at room temperature 300K. As the ambient temperature increases from 250K to 400K, the V_{th} shift is more pronounced in the acceptor trap case (Fig.5a-b), whereas the donor trap causes higher SS (Fig.5c-d). In addition, the concentration of the trap charges has a significant role as it alters the overall channel charge. Thus, the peak concentration of the trap charges varies from 10^{11} to 10^{13} cm^{-2} (Fig.6a-f), keeping the location fixed, i.e., donor at CB and acceptor trap at VB edges, respectively. With an increase in acceptor trap concentration, the OFF current improves significantly, which results in a promising improvement in the I_{ON}/I_{OFF} ratio, and threshold voltage. However, for the donor trap, the effective charge concentration in the channel region enhances significantly which deteriorate the overall NSFET electrical characteristics. Further, the combined impact on the NSFET performance will provide an optimal design guideline to detain the early aging of the device. Fig. 7(a-e) shows the trap and temperature-assisted reliability analysis of the NSFET, showing the acceptor trap charges are more susceptible to the EOL variation as compared to the donor trap charge.

Conclusions: In this work, using well-calibrated TCAD models, we comprehensively analyzed the impact of trap charges and SHE-induced thermal degradation in Nanosheet FET. Both the traps, i.e., donor and acceptor, have been considered separately and investigated the effect of their location and concentration. When the ambient temperature varies under the circumstances of interface trap charges, the impact becomes more severe. Thusly, trap and temperature-induced V_{th} shifts occur, which results in end-of-life (EOL). Hence, a proper investigation of these performance barrier parameters is worthy to achieve optimal NSFET performances.

Acknowledgment: N. Bagga greatly acknowledge the support obtained from PDPM-IIITDMJ (project no. IIITDMJ/ODRSPC/2022/88)

References: [1] J.-C. Barbé etal., SISPAD 2017. [2] S. Natarajan et al., IEDM 2015. [3] S. Rathore et. al., SST 2022.[4] Q. Liu et. al. IEDM, 2013 [5] Sentaurus TCAD Manual. [6] N. Loubet et al., IEDM 2017.

Trap and Self-Heating Effect Based Reliability Analysis to Reveal Early Aging Effect in Nanosheet FET

Motivation:

(a) Trap concentration and location of the random traps plays a significant role in deciding the device characteristics.

(b) Ambient temperature is a key factor which decides the aging of the device.

(c) Self-heating Effect (SHE) might be a crucial issue for early aging of Nanosheet FET (NSFET).

Obj. 1: Impact of Donor and Acceptor traps and their location with respective conduction and valance bands, defining the barrier in performance of NSFET.

Obj. 2: Analyzing the impact of ambient temperature with SHE for fixed traps at fixed location.

Obj. 3: Impact of variation in trap concentration on device performance.

Obj. 4: Device Aging analysis, i.e., end of lifetime (EOL).

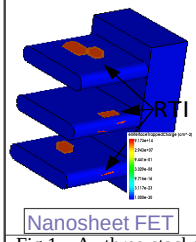


Fig.1. A three-stacked Nanosheet FET containing the random traps at the *Si-SiO*$_2$ interface.

Calibration of TCAD Setup

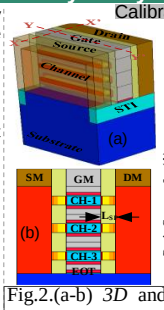
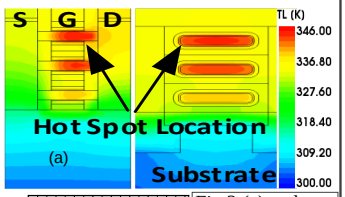


TABLE-I

Parameter	Value
Gate Length	12 nm
EOT	0.9 nm
Spacer length (L_{SP})	5 nm
Sheet Thickness (t_{CH})	5 nm
Sheet Width (W)	25 nm
Sheet Spacing	11 nm

Fig.2.(a-b) *3D* and cross-sectional view of the baseline Nanosheet FET used for TCAD calibration; (c) TCAD model calibrated I_{DS}-V_{GS} with Exp. data [6]; The parameters used in the simulation are mentioned in Table-I.

SHE in NSFET



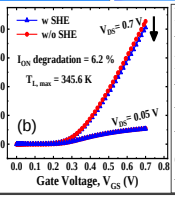


Fig.3:(a) shows the hotspot in NSFET, found at the channel near to the drain; (b) the performance degradation due to SHE.

Obj.1: Impact of Trap Energy Location

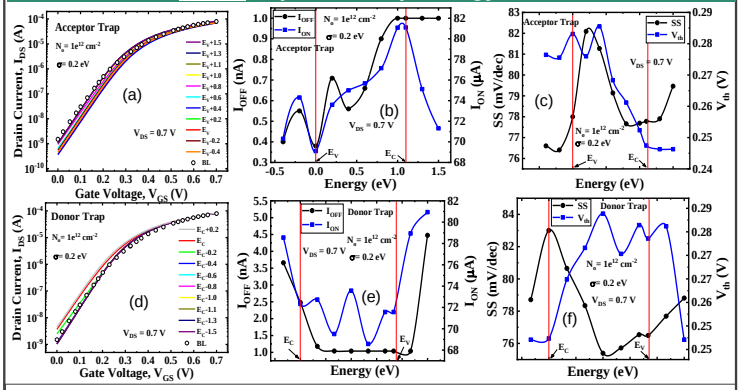


Fig.4. Impact of donor and acceptor trap location is investigated. For donor [acceptor] traps, the location is measured with respect to the conduction [valence] band in the interval of φ_F (i.e., $^{kT}/_{q}\ln(N_{ch}/n_i)$, where N_{ch} and n_i are channel and intrinsic doping. (a) shows that when acceptor trap location moves towards conduction band (CB), then threshold voltage (V_{th}) decreases as the charges does not get any vacant energy state near the valence band (VB); (d) location of donor traps near the CB, provides extra charges and thus V_{th} decreases. (b-c [e-f]) shows I_{ON}, I_{OFF}, SS and V_{th} variation when acceptor [donor] trap lies in energy gap (E_C-E_V).

Obj.2: Effect of Ambient Temp. on SHE

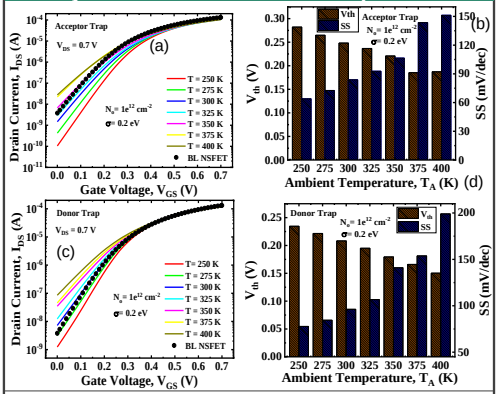


Fig.5. Consideration of SHE with varying ambient temperature effect for fixed trap concentration and location (acceptor [donor] traps taken at VB [CB] edges). (a-c) I_{DS}-V_{GS} characteristics showing that donor traps are prone to temperature effect as they provide extra charges in CB and thus increases the I_{OFF}; however, the impact is less when acceptor traps are considered. (c-d) change in V_{th} and SS with temperature.

Obj. 3: Impact of Trap Concentration

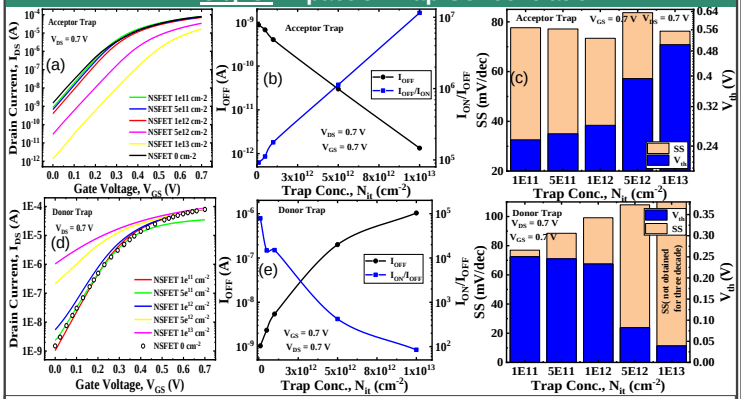


Fig.6. Impact of varying the peak trap concentration from $1x10^{11}$ cm^{-2} to $1x10^{13}$ cm^{-2} for acceptor [donor] trap with standard deviation (σ = 0.2eV). Increasing the acceptor trap concentration will increases the threshold voltage (V_{th}) (c), improves the NSFET device's performance (a-c). However, as donor trap concentration increases, the extra charges are accumulated in the active channel region which will decrease the V_{th} (f) and causes deteriorates the NSFET performances (d-f).

Obj. 4: Aging Analysis (EOL)

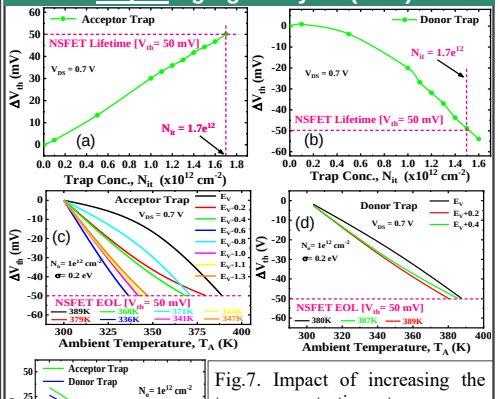


Fig.7. Impact of increasing the trap concentration, trap energy location, and trap assisted SHE deteriorates the EOL for donor [acceptor] trap. Acceptor traps are more susceptible to these variations.

A Physical Model for Long Term Data Retention Characteristics in 3D NAND Flash Memory

Rashmi Saikia and Souvik Mahapatra*
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, 400076, India
* Phone: (+91) – (222) – 572-0408, E-mail: souvik@ee.iitb.ac.in

***Abstract*— An Activated Barrier Double Well Thermionic Emission (ABDWT) model is used to simulate long-term data retention (DR) in 3D NAND Flash memory cells. Contribution due to only charge De-Trapping (DT) when adjacent cells are at the same charged state, and additional contribution due to charge Lateral Migration (LM) when adjacent cells are at different charged state are modeled. The Temperature (T), Program Level (PL) and Erase-Program Cycle (EPC) impacts are studied. Data from published sources are used to validate the model, and parameters are listed.**

Introduction: 3D-NAND Flash with Silicon Nitride (SiN) based Charge Trap Layer (CTL), Fig.1 [1]-[2], is the latest technology for solid-state storage products. Multiple charge storage levels are placed in a memory cell to achieve multi-bit capability, at the cost of reduced spacing between levels due to the fundamental limit of maximum charge storage in CTL. DR is a key reliability issue, resulting in charge loss from a given PL and shifting of the Cell V_T Distribution (CVD) towards lower PL (V_T: threshold voltage). This reduces the spacing between the lower CVD tail of a higher PL and upper CVD tail of the adjacent lower PL, resulting in Fail Bit Count (FBC) during memory read operation. DR is classified into short and long term modes [3]-[4]. Due to inherent limitation of long measurement time, long term DR is measured at an elevated T over a relatively shorter time (~ hours), and models are used to estimate the same at end-of-use (~ years) at lower (~ operation) T. Only empirical models are used for the same as of today, with uncertainties in the model parameters (*e.g.*, the use of different Arrhenius T activation energy (E_A) values for different PL) [3]-[5]. In this work, a physics-based ABDWT model [6] is invoked to do the same, with consistent values of model parameters.

Overview of DR mechanisms: Charges stored in CTL during the programming phase can reduce in time due to multiple processes, Fig.2 [5]: (1) the charges can de-trap (DT), (2) due to continuous CTL throughout the bit-line, the charges can migrate between the adjacent cells when they are not at same PL (LM), (3) generation of traps in the Tunnel Oxide (TO) after multiple EPC can result in charge loss via Trap-Assisted-Tunneling (TAT), although this mainly affects CVD tails and more so after higher number of EPC steps, and (4) Vertical Redistribution (VR) of charges inside CTL of a cell, mainly impacting short-term DR. This work focuses on long-term DR (so VR is ignored) due to DT and LM (TAT is not considered as only median of the CVD is modeled).

Model details: ABDWT model, Fig.3 [6], is used for calculating the time kinetics of changes in cell V_T (ΔV_T) due to DT and LM (handled separately, total ΔV_T is the sum of individual models). The model has two stable energy states, E_1 and E_2, separated by an energy barrier E_B. The barrier has a Gaussian distribution in energy, with T activated mean (E_{B_MEAN}) and spread (E_{B_SPREAD}) with activation energy E_{A_MEAN} and E_{A_SPREAD} respectively. Once the cell is programmed, charges in the CTL are considered to be in the E_1 state. The charges in CTL set up an electric field (E_{OX}) across TO, blocking oxide (BO), and between adjacent cells in the bit-line when their PL is different. DT is caused by E_{OX} across TO (DT across BO is ignored due to its larger thickness), and LM is due to E_{OX} between adjacent cells. E_{OX} lowers E_B (by Δ) and E_2 (by $m*\Delta$), $\Delta = \gamma*E_{OX}$ and γ is the field acceleration factor. The lowering of E_B triggers over-the-barrier thermionic emission and charge transfer from state E_1 to state E_2; reduction in charges for state E_1 is related to DR. N_0 is a pre-factor related to initial trap density, β is thermal energy (1/kT), and υ is the attempt-to-escape frequency. E_1 is the reference level, all energies are w.r.t E_1.

Model validation: Fig.4 and Fig.5 show experimental ΔV_T from [7], at 2 different DR bake T, and for 3 different DR conditions at each T: the target cell is fixed at 4V, while adjacent cells along the bit-line are at 3V (PPP), 0V (NPN), and –1.5V (EPE), along with model calculation, with underlying DT and LM components shown in Fig.6. In PPP, the LM has least contribution and DT due to E_{OX} across TO dominates. As the neighboring cell V_T reduces to N and E, LM due to E_{OX} between cells increases but DT stays fixed. Both components get activated at higher T. Parameters are listed in Table-I, identical values are used for both DT and LM to model different experimental conditions. Fig.7 and Fig. 8 show experimental ΔV_T from [8], for solid (PPP) and checkered (EPE) patterns, with target cell at different PL (P2 to P7), along with model calculation. In this case, DR experiments are done after the cells are subjected to 300 EPC. LM contribution is negligible for solid pattern, and it gets triggered in checkered pattern and shown in Fig.9. The DT contribution increases for higher PL, however for a particular PL, it is kept fixed between solid and checkered patterns (*i.e.*, the DT model of Fig.7 is also used for Fig.8). The LM contribution increases for higher PL. Parameters are listed in Table-II. Except E_2, identical values are used for all parameters to model DT and LM for different PL. We speculate that the E_2 value is different between DT and LM due to charge accumulated in the intercell regions during EPC in these experiments. Some of the parameters are slightly different between the two sources of data, and is presumably due to structural and material properties being different between different production lines.

Conclusion: Long-term DR is successfully modeled with DT and LM contributions calculated using the ABDWT framework. DR gets accelerated at higher T, from higher PL, and for checkered (compared to solid) programming pattern, with more difference in PL between the adjacent cells, and all aspects can be handled by ABDWT. Importantly, only DT affects DR in the solid pattern configuration, and identical DT contribution is used for a given PL between solid and checkered patterns. Therefore, this physics-based framework can be used to determine end-of-life DR during the product qualification phase.

References: [1] Micheloni, Proc IEEE, 2017, P. 1609 [2] Raghunathan, EDTM, 2020, P.1 [3] Woo, VLSI, 2019, P.T214 [4] Kim, TED, 2020 P. 5472 [5] Woo, IRPS, 2020, P.1 [6] Choudhury, JEDS, 2020, P. 1281 [7] Kang, VLSI, 2015, P.T182, [8] Mizoguchi, IEDM, 2017, P.19.2.1

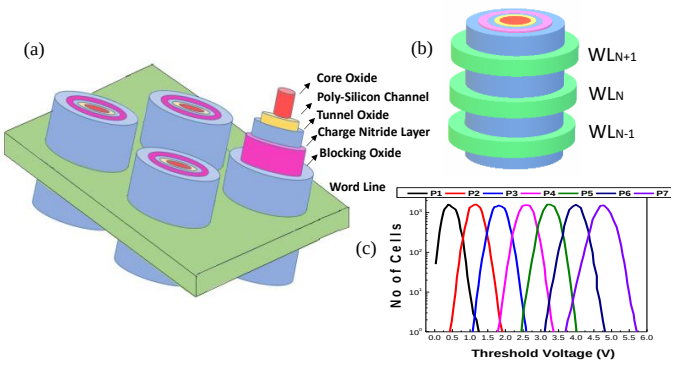


Fig.1. (a) Schematic of BiCS memory cell with the highlighted layers of core oxide, poly-silicon channel, tunnel oxide, charge trap nitride layer, blocking oxide and the metal gate, (b) Schematic showing the stacked word-lines (WL_N) and (c) CVD of a triple level cell (TLC) flash from state P1 – P7.

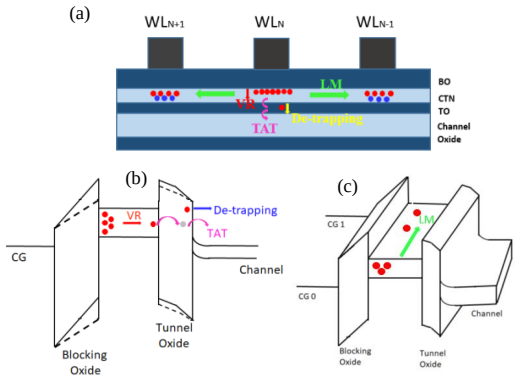


Fig.2. (a) Cross sectional explanation of VR, LM, de-trapping and TAT mechanism. Schematic of band diagram of charge loss during short term retention, (b) vertical charge loss (solid line-before, dash line-after), (c) lateral charge loss. Re-created as per [3].

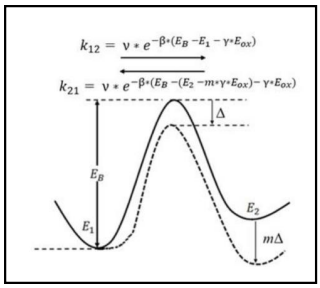


Fig.3. Schematic of ABDWT model with the rate equations.

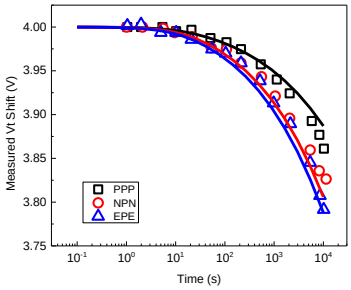


Fig.4. Modeling of Δ V_T shifts versus retention time under different adjacent cell states (neutral (N), erase (E) and program (P)) at 90°C [7].

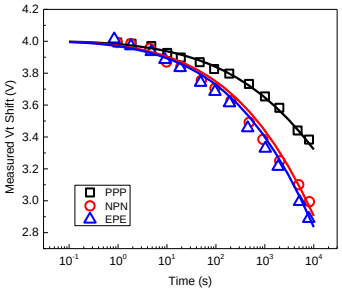


Fig.5. Modeling of Δ V_T shifts versus retention time under different adjacent cell states (neutral (N), erase (E) and program (P)) at 200°C [7].

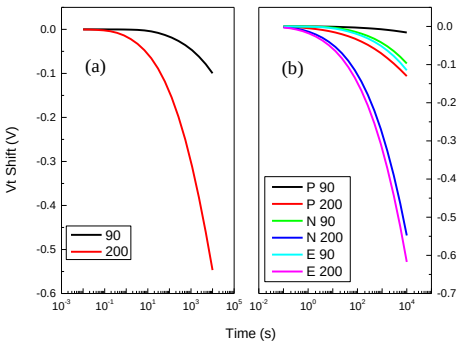


Fig.6. Component time kinetics (a) de-trapping and (b) LM component at 90° C and 200° C for PPP, NPN and, EPE condition.

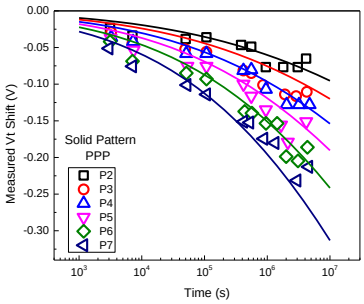


Fig.7. Modeling of the Δ V_T shift during solid patterns (PPP) at 85°C with cell state changing from P2 state to P7 state [8].

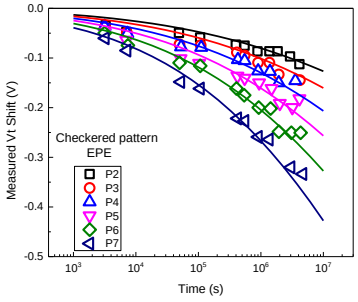


Fig.8. Modeling of the Δ V_T shift in checkered patterns (EPE) at 85°C for target cell state changing from P2 state to P7 state [8].

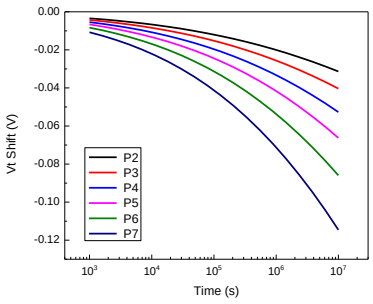


Fig.9. Time kinetics of LM component during checkered condition for P2 state to P7 state.

Model Parameters

DT

E2=0.12, E_{B_MEAN}=2.8, E_{B_SPREAD}=0.3, m=8, E_{A1}=0.013, E_{A2}=0.0047, γ=8.5e-9, N_0=2e13

LM

E2=0.12, E_{B_MEAN}=2.8, E_{B_SPREAD}=0.3, m=8, E_{A1}=0.013, E_{A2}=0.0047, γ=8.5e-9, N_0=2e13

Table I. ABDWT model parameters for de-trapping and lateral migration for Fig 4 and Fig 5.

Model Parameters

DT

E2=0.12, E_{B_MEAN}=2.8, E_{B_SPREAD}=0.3, m=7, E_{A1}=0.013, E_{A2}=0.0047, γ=1.5e-9, N_0=4e13

LM

E2=0.16, E_{B_MEAN}=2.8, E_{B_SPREAD}=0.3, m=7, E_{A1}=0.013, E_{A2}=0.0047, γ=1.5e-9, N_0=4e13

Table II. ABDWT model parameters for de-trapping and lateral migration for Fig 7 and Fig 8.

An Atomistic Modelling Framework for Valence Change Memory Cells

M. Kaniselvan, M. Luisier, and M. Mladenovic
Integrated Systems Laboratory, ETH Zurich. Email: mkaniselvan@iis.ee.ethz.ch

Introduction - Neuromorphic computing units require the development of solid-state synapses that are often realized in the form of devices with adjustable conductance. Amongst such devices are valence change memory (VCM) cells. These are two-terminal metal-oxide-metal stacks across which applied voltages can drive the creation of oxygen vacancies and their redistribution into a conductive filament. The resulting conductance can then be measured with a low readout voltage. VCMs are both relatively simple to fabricate and exhibit an especially large dynamic range.

Simulation of these devices is, however, complicated by the stochasticity of their operation and the atomistic granularity of their conductance. Electrical currents can be treated with analytical trap-assisted tunneling models [1–3], but the required equations are typically derived for single trap energies, and may not fully describe transport through the inhomogenous defect distribution in a VCM. Simulating both atomic rearrangement and realistic transport properties therefore necessitates a more fundamental level of theory.

Here we present a framework dedicated to modelling VCM. Our method relies on a stochastic Kinetic Monte Carlo (KMC) model parameterized with density functional theory (DFT), followed by *ab initio* quantum transport simulations, all preformed on the same atomic grid. It thus captures the growth and dissolution of oxygen vacancy filaments through the VCM cell and the electrical current that flows through them.

Model - **Fig. 1a** shows a schematic of the nominal VCM cell considered: an amorphous HfO_2 oxide with TiN electrodes. The oxide is generated by subjecting a block of monoclinic HfO_2 to a melt-cool-anneal process [4], and subsequently relaxing the atomic positions using the cp2k code [5]. TiN electrodes are then attached along the transport direction.

Oxygen vacancy (V_O^{+2}) and ion (O^{-2}) rearrangements under an applied voltage (V_{app}) are modelled with an in-house KMC code. In this method, an event is selected once per time step from the set of every vacancy/ion pair generation, recombination, and diffusion process in the oxide. The selection probability $P^x_{i,j}$ of event 'x' between sites 'i' and 'j' is

$$P^x_{i,j} \propto \exp\left(-\frac{E^x_A - E_{i,j}}{k_B T}\right), \tag{1}$$

where E^x_A is the zero-field activation energy of event x. These energies are found for amorphous HfO_2 through DFT calculations. We assume a reduced activation energy for vacancy generation at the active (top) electrode [6]. Subtracted from this is the energy provided by the applied field $E_{i,j} = q \cdot (\phi_i - \phi_j)$, considering the charge (q) and potential (ϕ) for each pair of sites involved. Clustered vacancies are presumed conductive and first set to a charge state of zero (V_O^0); ϕ at each site is then determined by treating the oxide as a network of nodes across which V_{app} dissipates. Events occur until the simulation timescale reaches the intended duration for which V_{app} is applied. A device snapshot is then generated. **Fig. 1b** presents an overview of this model.

To calculate the conductance of each snapshot, we use the quantum transmitting boundary method as implemented in the OMEN code [7, 8]. Coherent transport occurs through states $\psi(E)$, which are found by solving:

$$(E \cdot S_{cp2k} - H_{cp2k} - \Sigma^{RB}(E)) \cdot \psi(E) = Inj(E) \tag{2}$$

Here, Inj(E) describes carrier injection from the contacts, which are coupled to the device through $\Sigma^{RB}(E)$. H_{cp2k} and S_{cp2k} are the Hamiltonian and overlap matrices, respectively, as produced with cp2k for each of the KMC snapshots. Due to the localized nature of the underlying Gaussian-type orbitals, H_{cp2k} and S_{cp2k} have a block-structure representing atomic layers as pictured in **Fig. 1a**. The O^{-2} ions are not included in electronic structure calculations.

Results - We apply this model to simulate the operation of a TiN/HfO_2/TiN VCM cell with a cross-section of 1.67 x 1.63 nm^2, for a total of 2252 atoms. Starting from a structure with a formed filament (**Fig. 2a**, left), a V_{app} of -2 V is applied to switch to the high resistance state (HRS, **Fig. 2a**, middle). V_{app} = 5 V is then applied to the HRS to recover a low resistance state (LRS, **Fig. 2a**, right). In each case, the duration of V_{app} is 10 ms. The switching process relies on the recombination and generation of V_O^{+2} near the active electrode; recombination caused by the reverse bias creates a tunneling gap in the filament, lowering the transmission function and thus the device conductance. A transition back to the LRS occurs when a sufficiently high V_{app} regenerates vacancy/ion pairs in this gap, aided by the steeper local electric potential in this case (**Fig. 2b**). The difference in transmission between these two states (**Fig. 2c**) is highest near the conduction band of HfO_2, consistent with the location of V_O^{+2} vacancy defect states at this energy range [9]. **Fig. 3** plots a full switching cycle, showing both the current (**Fig. 3a**) and conductance (**Fig. 3b**) at each intermediate V_{app}. The device has a current hysteresis typical of VCM, with a conductance ratio of ∼10. The asymmetry in the V_{app} between the HRS and the LRS transitions stems from the length of the conductive filament being lower in the initial device than in the final LRS, (see **Fig. 2a**), and from the recombination of V_O^{+2}/O^{-2} being far more energetically favorable than their generation.

Conclusion - We combined KMC and *ab initio* quantum transport on an atomic lattice to model resistive switching in a VCM cell. This framework will provide insight towards the optimization of VCM material stacks and geometries.

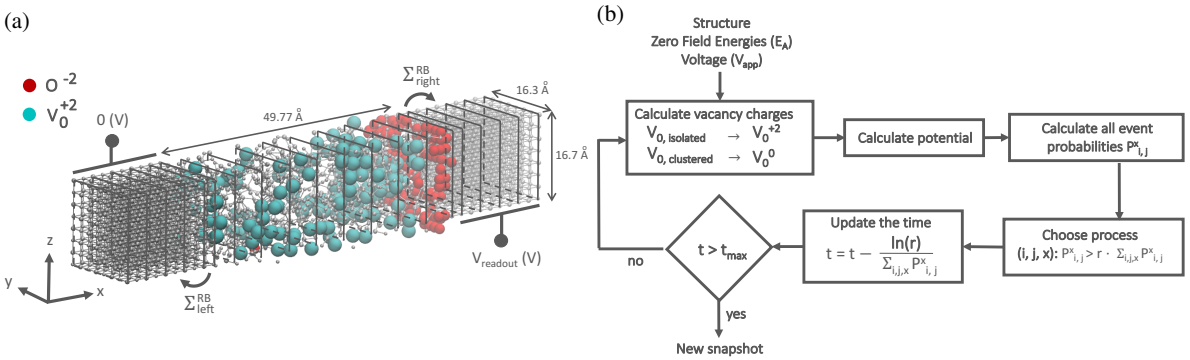


Fig. 1: Details on the approach used to model the conductance through a TiN/HfO_2/TiN VCM cell under an applied voltage. **(a)** A typical metal-oxide-metal structure. The device is partitioned into 'blocks' of atoms representing the underlying block-structure of the Hamiltonian (H_{cp2k}) and overlap (S_{cp2k}) matrices in **Eq. 2**. **(b)** Flowchart of the developed Kinetic Monte Carlo model which determines the state of atomic rearrangement under V_{app}. The final snapshot of the device is generated at time t_{max}. r represents a random number. $P^x_{i,j}$ is calculated according to **Eq. 1**.

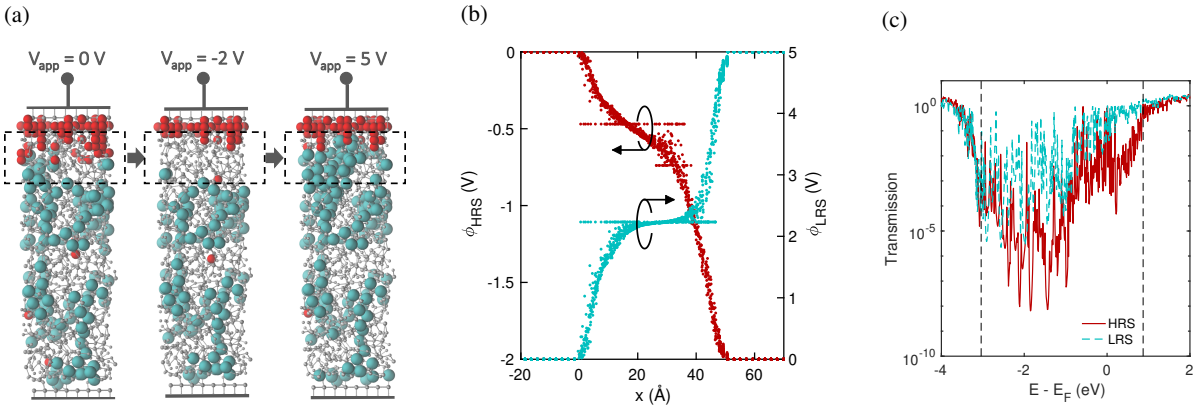


Fig. 2: Properties of the high (HRS) and low resistance state (LRS) states of the TiN/HfO_2/TiN VCM cell during a single switching cycle. In **(a)**, V_O^{+2} and O^{-2} positions along the oxide are shown for (left) the formed filament, (middle) the HRS, and (right) the LRS. The areas near the active electrode, where the filament length varies most with the V_{app}, are indicated with dashed boxes. **(b)** Electrostatic potential (ϕ) along the oxide for the HRS (red) and LRS (blue) from (a). The HfO_2 oxide spans from 0 Å to 50 Å. **(c)** Transmission through the HRS and LRS from (a). The dashed lines indicate the estimated bandgap of the HfO_2 layer.

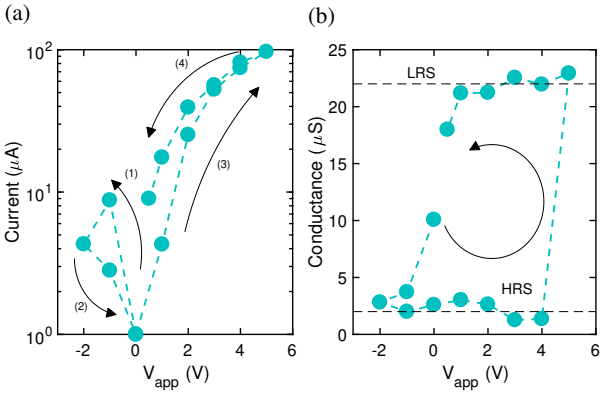


Fig. 3: Current **(a)** and conductance **(b)** of the device shown in **Fig. 1a** during a switching cycle, from the initial filament (0 V) to the HRS (-2 V) and finally to the LRS (5 V).

References

(1) Padovani, A.; Larcher, L.; Woo, J.; Hwang, H. In *2017 17th Non-Volatile Memory Technology Symposium (NVMTS)*, IEEE: 2017.

(2) Bersuker, G.; Gilmer, D. C.; Veksler, D.; Kirsch, P.; Vandelli, L.; Padovani, A.; Larcher, L.; McKenna, K.; Shluger, A.; Iglesias, V.; Porti, M.; Nafria, M. *Journal of Applied Physics* **2011**, *110*, 124518.

(3) Kopperberg, N.; Wiefels, S.; Liberda, S.; Waser, R.; Menzel, S. *ACS Applied Materials & Interfaces* **2021**, *13*, 58066–58075.

(4) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. *Comp. Phys. Comm.* **2022**, *271*, 108171.

(5) Kühne, T. D. et al. *The Journal of Chemical Physics* **2020**, *152*, 194103.

(6) Traore, B.; Blaise, P.; Sklenard, B.; Vianello, E.; Magyari-Kope, B.; Nishi, Y. *IEEE Transactions on Electron Devices* **2018**, *65*, 507–513.

(7) Luisier, M.; Schenk, A.; Fichtner, W.; Klimeck, G. *Physical Review B* **2006**, *74*.

(8) Ducry, F.; Aeschlimann, J.; Luisier, M. *Nanoscale Advances* **2020**, *2*, 2648–2667.

(9) Robertson, J. *Reports on Progress in Physics* **2005**, *69*, 327–396.

An inner gate as enabler for vertical pitch scaling in macaroni channel gate-all-around 3-D NAND flash memory

D. Verreck, A. Arreghini, G. Van den bosch and M. Rosmeulen
imec, 3001 Leuven, Belgium, email: devin.verreck@imec.be

Scaling vertical cell pitch to increase bit density in 3-D NAND flash memories degrades both the cell transistor characteristics and the memory operation. Here, we therefore investigate an inner gate to mitigate the scaling impact in macaroni channel devices. We evaluate several scenarios with varying complexity using calibrated TCAD simulations: from keeping the inner gate voltage grounded to coupling it to the read gate. We find a trade-off between improved cell transfer characteristics and program voltage determined by the inner gate to read gate coupling ratio.

NAND flash memories are currently in an era of "happy scaling" thanks to the transition from 2-D arrays to 3-D vertical strings [1]. In a vertical architecture, all cells on a string are fabricated at the same time by the deposition of a layered stack, followed by memory hole etch and filling steps. The number of cells on a string, and thus bit density, can be increased by adding layers to the stack, while keeping the vertical cell pitch relaxed to avoid short channel effects. As the memhole aspect ratio grows, however, the etch becomes so challenging that vertical pitch scaling is again required to limit the total stack height [2]. Unfortunately, the loss of gate control associated with the reduction of the cell dimensions results in undesirable effects: an increasingly negative cell threshold voltage (V_{TH}) and an increase in programming voltage (V_{PGM}) [3]. In planar structures, a back gate has been proposed to mitigate these scaling effects [4]. Here, we study whether a similar principle can be applied to 3-D NAND macaroni channel strings by replacing the inner oxide with an inner gate (IG). We use calibrated TCAD simulations to assess the joint impact on cell characteristics and V_{PGM} for various operation schemes.

The three-gate structure under study is shown in Fig.1(a) with a cross-section of the memhole in Fig.1(b) and simulation parameters in Fig.1(c). This structure mimics our experimental test vehicles, with the addition of an inner gate contact and dielectric in the center of the memhole. The inner gate covers the entire backside of all three gates. The vertical pitch is scaled by varying the gate lengths (L_G) and intergate spacings (L_{IGS}).

The simulations are performed with Global TCAD Solutions software [5], which incorporates a jointly developed charge trap layer (CTL) memory operation model [3]. During program, carriers are injected with a Wentzel-Kramers-Brillouin tunneling approach and then distributed over the CTL according to a Gaussian profile. This model has been calibrated to our in-house experimental data [3]. For the read operation, a standard drift-diffusion model is employed in the channel layer. Since we are interested in relative improvements, we do not consider the channel grains here.

We first simulate the scaling behavior for the reference case without an IG, showing a strongly negative V_{TH} shift for the smallest pitch dimensions (Fig.2(a)), combined with an increased V_{PGM} (Fig.2(b)) and decreased on-current (I_{ON}) (Fig.2(c)). Fig.3(a) shows that the center gate (CG) loses channel control when the side gates are closer, resulting in a reduced energy barrier in the subthreshold regime (Fig.3(b)). This effect, which we refer to as neighbor-induced barrier lowering (NIBL), has a relatively larger impact at shorter L_G.

Next, we assume the IG grounded to 0V. This is the most straightforward case to implement, e.g. through a connection with the source contact. We find that a V_{IG} of 0V for an EOT_{IG} symmetrical to the ONO EOT of 11.5nm has small positive impact on the cell V_{TH} relative to the reference case (Fig.4(a)), while V_{PGM} is almost unaffected (Fig.4(b)). I_{ON} is slightly degraded for the larger pitches (Fig.4(c)), as the string resistance increases due to depletion by the IG. Reducing EOT_{IG} to 5nm increases the positive ΔV_{TH} (Fig.5(a)), while keeping V_{PGM} and I_{ON} impact limited (Fig.5(b-c)). The band diagram along the CG center in Fig.6 explains that V_{PGM} is insensitive to V_{IG} because the channel screens the IG potential, leaving the electric field over the ONO stack quasi unaffected. This remains true even for negative V_{IG}. It also means the IG cannot be used to improve programming behavior, and we therefore keep V_{IG}=0V during program for subsequent cases.

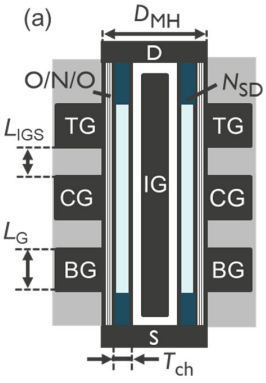
In the next scenario, we connect IG to CG during the read operation, which results in a significant improvement of V_{TH} and I_{ON} (Fig.7(a) and (c)). The improvement in V_{TH} goes up to 4V for the most scaled cases. This is thanks to the double gate action of IG and CG. V_{PGM} is strongly increased, however (Fig.7(b)). During the read operation, the channel is now turned on by both CG and IG and since charge is only programmed in the CTL on one side of the channel, its control on V_{TH} is strongly reduced compared to the reference case without the IG.

Finally, we couple IG to CG with a varying ratio (V_{IG}=V_{CG}/D), which results in a trade-off between V_{TH} and I_{ON} on one hand and V_{PGM} on the other (Fig.8(a-c)). The results interpolate between the V_{IG}=0V and coupled V_{IG}=V_{CG} cases discussed above. Depending on the application, a compromise can be found, e.g. for D=2, a 59% V_{TH} improvement is achieved at the cost of a 17% V_{PGM} increase.

We conclude that an inner gate can be used to mitigate negative V_{TH} shifts and improve I_{ON} in vertically scaled 3-D NAND cells at the expense of required programming voltage. We found that the trade-off depends on the coupling ratio of the IG to the read gate. We showed that an IG has very limited impact on the program operation due to channel screening.

This work was supported by imec's Industrial Affiliation Program for storage memories.

[1] Alsmeier et al., IEEE Int. Elec. Dev. Meeting., 2020.
[2] Yanagihara et al., IEEE IMW, 2012.
[3] Verreck et al., IEEE Int. Elec. Dev. Meeting, 2021.
[4] Lin et al., IEEE Trans. Elec. Dev. 58, 11, 2011.
[5] http://www.globaltcad.com/



(c)

D_{MH}	120nm
O/N/O	6/6/6nm
N_{SD}	1e20cm^{-3}
T_{ch}	10nm
V_{SEL}/V_{DS}	7/0.1V
$L_{G/IGS}$	10,15,20, 30,40nm

Fig. 1. (a) Simulated macaroni structure with top, center and bottom gate (TG,CG,BG), (b) memhole cross-section and (c) simulation parameters.

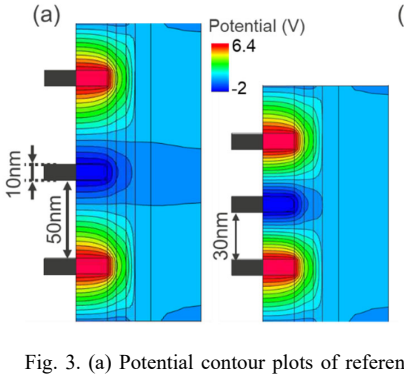


Fig. 3. (a) Potential contour plots of reference case without inner gate for varying L_{IGS} at fixed L_G. (b) Conduction band profile along the length of the channel at 1nm from the tunnel oxide interface.

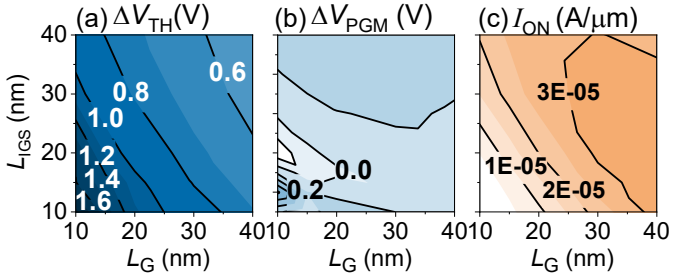


Fig. 5. L_G, L_{IGS} scaling for device with IG with V_{IG}=0V and EOT$_{IG}$=5nm. (a) V_{TH} difference of CG relative to non-IG case (b) V_{PGM} difference relative to non-IG case and (c) device current at an overdrive of V_{TH}+2V.

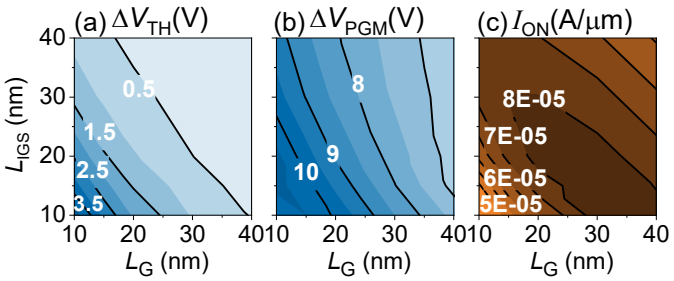


Fig. 7. L_G, L_{IGS} scaling for device with IG with V_{IG}=V_{CG} during read and EOT$_{IG}$=11.5nm. (a) V_{TH} difference of CG relative to non-IG case (b) V_{PGM} difference relative to non-IG case and (c) device current at an overdrive of V_{TH}+2V.

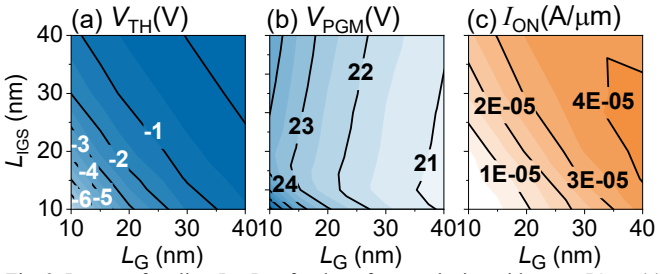


Fig. 2. Impact of scaling L_G, L_{IGS} for the reference device without an IG on (a) cell V_{TH} of CG extracted at I_{OFF} of 1e-9A/μm, (b) V_{PGM} required to reach ΔV_{TH} of 5V on CG and (c) device current at an overdrive of V_{TH}+2V.

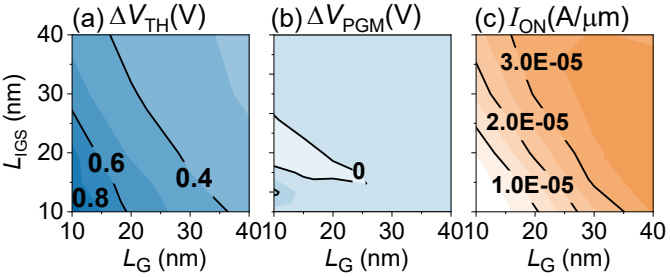


Fig. 4. L_G, L_{IGS} scaling for device with IG with V_{IG}=0V and EOT$_{IG}$=11.5nm. (a) V_{TH} difference of CG relative to non-IG case (b) V_{PGM} difference relative to non-IG case and (c) device current at an overdrive of V_{TH}+2V.

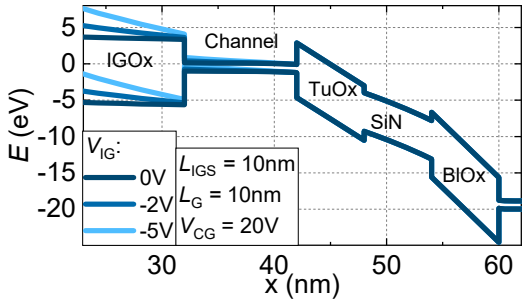


Fig. 6. Energy band diagram along a cutline through the center of CG for varying V_{IG} during program operation. x=0nm corresponds to the memhole center.

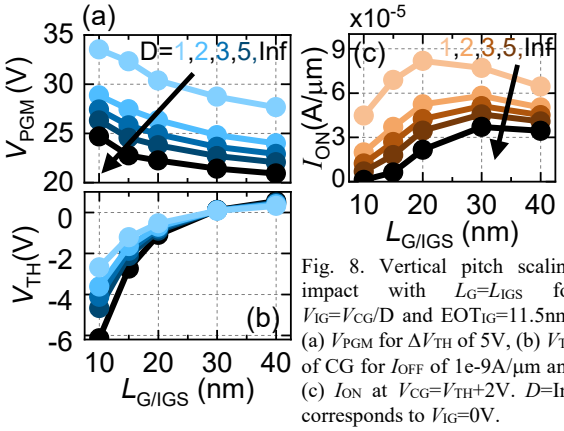


Fig. 8. Vertical pitch scaling impact with L_G=L_{IGS} for V_{IG}=V_{CG}/D and EOT$_{IG}$=11.5nm. (a) V_{PGM} for ΔV_{TH} of 5V, (b) V_{TH} of CG for I_{OFF} of 1e-9A/μm and (c) I_{ON} at V_{CG}=V_{TH}+2V. D=Inf corresponds to V_{IG}=0V.

Insights into Few-Atom Conductive Bridging Random Access Memory Cells with a Combined Force-Field / *ab initio* Scheme

J. Aeschlimann, M. H. Bani-Hashemian, F. Ducry, A. Emboras, and M. Luisier
Integrated Systems Laboratory, ETH Zürich, Switzerland, e-mail: aejan@iis.ee.ethz.ch

Introduction: Conductive bridging random access memory (CBRAM) cells are composed of a metal/oxide/metal structure where the insulating layer can be switched between a high- and low-resistance state through the reversible growth of a metallic filament upon application of an external potential (Fig. 1a). To drive this technology towards its limits, the device sizes have been scaled down such that only few atoms are involved in their ON/OFF switching processes [1]. To increase the device performance and reliability, a clear understanding of the switching mechanism is needed at the atomic level. Molecular dynamics based models have been suggested to investigate the metal migration through the oxide [2]. Such approaches typically include a subsequent percolation analysis to identify the ON- and OFF-states.

In this work, we present a model for Ag/a-SiO_2 cells based on force-field molecular dynamics (FFMD). In contrast to *ab initio* MD (AIMD) [3], it is possible to simulate time lengths of several nanoseconds with FFMD, covering full SET and RESET processes. To determine the resistance state of the obtained filamentary structures, we then perform *ab initio* quantum transport (QT) calculations. Through this scheme, we benefit from both the computational efficiency of FF-based methods and the accuracy of *ab initio* calculations. With a structural analysis of the oxide, we identify preferred channels containing wide SiO_2 rings through which Ag^+ ions migrate during the switching process. We demonstrate that moving only few atoms in such channels can change the resistance state by several orders of magnitude.

Approach: Ag/a-SiO_2 cells have been created by stacking two Ag electrodes with a layer of amorphous SiO_2 (a-SiO_2) in between. The latter is generated by the melt-and-quench method [4]. Only samples free from structural defects were selected to avoid negative impacts (e.g., Fermi level pinning) on the QT calculations. A truncated cone-shaped filament made of 50 Ag atoms was inserted into the oxide as a seed, leaving a gap of 1.2 nm between the electrodes (Fig. 1b).

The FFMD simulations have been performed using QuantumATK S-2021.06 [5] and the moment tensor potential (MTP) framework [6]. The parameter file was trained using a set of representative AIMD trajectories performed with CP2K [7]. The influence of the external electric potential has been included by modifying the forces with an additional term $\overrightarrow{F} = q \cdot \overrightarrow{E}$, where q and E are the atomic charge and the applied electric field, respectively. The electric field is evaluated by a generalized Poisson solver [8] treating the various filament shapes accurately. The charges on the Ag atoms are assigned based on their coordination number.

For selected structures obtained by FFMD, *ab initio* QT calculations have been performed to extract their resistance state. To do so, the structures were annealed for 0.1 ps with AIMD. As a next step, the Hamiltonian and overlap matrices were prepared in CP2K using contracted Gaussian-type DZVP orbitals as basis set. They were finally passed to OMEN [9], a quantum transport device simulator.

Results: To verify that the low-resistance state is not limited by intrinsic defects in the oxide, we first performed QT calculations of pristine oxide structures with 3 different lengths. The conductance decreases exponentially with longer oxides, indicating that the current is dominated by tunneling effects and not by structural defect states (Fig. 1c).

FFMD simulations under an applied voltage of +7 V show the growth of a filament consisting of a single atom chain towards the counter electrode. After 6.5 ns, the device has reached its ON-state, exhibiting an ON/OFF resistance ratio of 4 orders of magnitude with a bridging filament containing as few as 5 atoms (Fig. 2). Fig. 3a confirms the conductive nature of the filament. The ON-state is kept during 2 ns without any applied voltage before the filament is disrupted by applying -1 V, which pushes the bridging atoms back.

A ring analysis of the SiO_2 reveals that the filament growth path depends on the local silica ring distribution (Fig. 3b). Similarly to the observation from diffusion studies [10], Ag^+ ions require more energy to pass through tighter rings than looser ones. Hence, the filament particularly forms in regions of many silica rings containing at least 6 Si atoms, whereas migration through 5-fold rings could not be observed (Fig. 3c). Hence, a filament can only become a bridging one if it finds a continuing channel of locally sparse SiO_2 regions.

Conclusions: We have developed a model to simulate the growth and dissolution process of bridging filaments in Ag/a-SiO_2 CBRAM cells using FFMD. Based on a SiO_2 ring analysis, we have identified specific paths through which Ag^+ ions preferably migrate. The resistance states have been evaluated through *ab initio* QT methods, pointing out that the movement of only few atoms can lead to changes of several orders of magnitude.

References: [1] B. Cheng et al., *Nature Comm. Phys.* 2, 28 (2019). [2] N. Onofrio et al., *Nature Materials* 14, 440 (2015). [3] J. Akola et al., *Phys. Rev. Mat.* 6, 035001 (2022). [4] F. Ducry et al., *Nanoscale Adv.* 2, 2648 (2020). [5] J. Schneider et al., *Modelling Simul. Mater.* 25, 085007 (2017). [6] A. V. Shapeev, *Multiscale Model. Simul.* 14, 1153 (2016). [7] T. D. Kühne et al., *J. Chem. Phys.* 152, 19 (2020). [8] M. H. Bani-Hashemian et al., *J. Chem. Phys.* 144, 044113 (2016). [9] M. Luisier et al., *Phys. Rev. B* 74, 205323 (2006). [10] K. Patel et al., *Microel. Reliability* 98, 144 (2019).

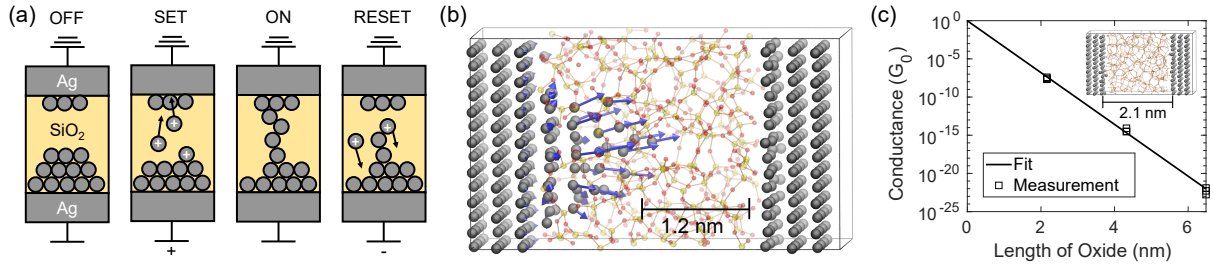


Fig. 1: (a) Illustration of the switching process occurring in CBRAM cells. The cell is built of two Ag electrodes and an insulating a-SiO_2 switching layer in between through which an Ag filament grows and dissolves, depending on the applied voltage. (b) Schematic view of the investigated Ag/a-SiO_2/Ag CBRAM structure. The Ag, Si, and O atoms are displayed as gray, yellow, and red spheres, respectively. A truncated cone-shaped Ag filament is inserted into the oxide as a seed, leaving a gap of 1.2 nm between the two electrodes. The cross section measures 2.06 x 2.04 nm^2. The blue arrows represent the local electric field the Ag filament atoms are exposed to when an external potential of +7 V is applied. (c) Electrical conductance (in conductance quantum G_0) as a function of the length of the oxide in pristine Ag/a-SiO_2/Ag CBRAM cells. 3 different samples have been considered with 3 different oxide lengths each. The exponential relationship indicates that the current is dominated by quantum tunneling and not by an intrinsic defect in the oxide. The inset shows the pristine structure with a gap of 2.1 nm.

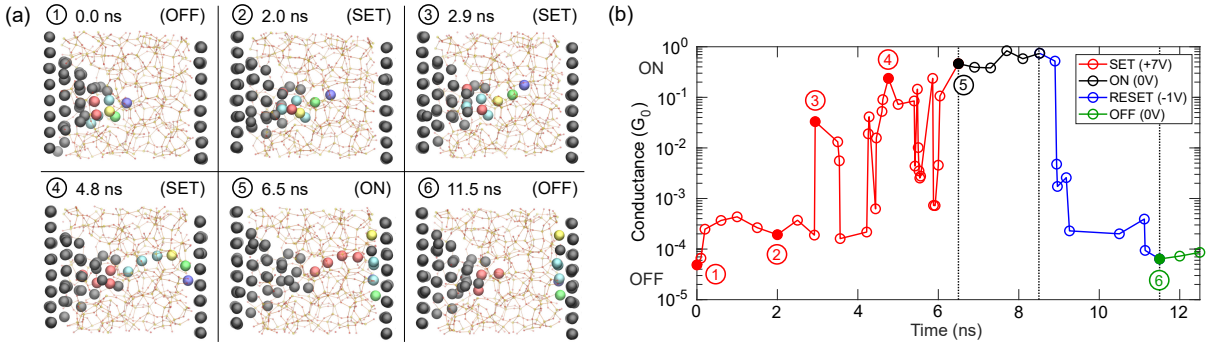


Fig. 2: (a) Evolution of the bridging filament in an Ag/a-SiO_2 CBRAM structure as a function of time. The structure displayed in Fig. 1b is used as starting point. Ag atoms are displayed as gray spheres, those contributing to the filament are colored. The a-SiO_2 matrix is shown in orange. To simulate the SET process, FFMD was performed at 300 K for 6.5 ns under an applied voltage of +7 V. A filament consisting of atoms originally located at the tip of the truncated cone grows into the oxide as a chain of single atoms. When the applied voltage is set to -1 V, the filament ruptures and the Ag atoms forming the filament (colored red) move backwards. (b) Time evolution of the conductance state. After an unstable phase, the filament finally reaches the ON-state after 6.5 ns, which is kept during 2 ns. The RESET to the OFF-state is achieved in less than 0.5 ns. The ON/OFF resistance ratio of the device is equal to 10^4.

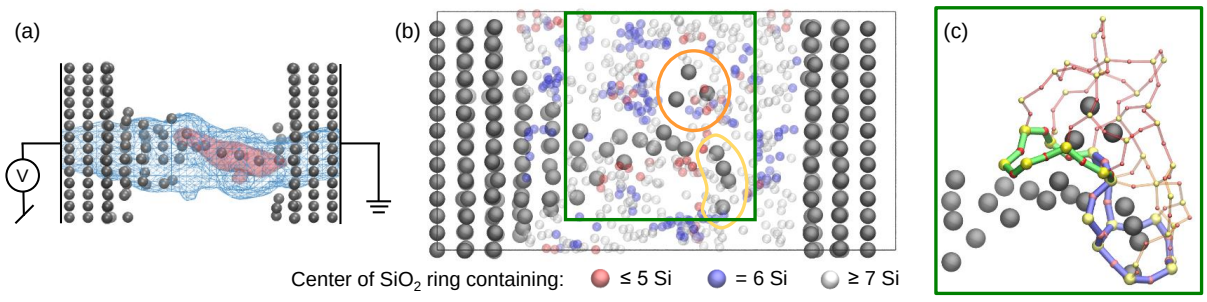


Fig. 3: (a) Ballistic current (illustrated as two different isosurfaces) through the structure representing the ON-state displayed in Fig. 2a. An external read voltage of V=0.1 V was applied. The maximum current is measured in the single atom chain of the filament. (b) Illustration of the filament growth in an Ag/a-SiO_2 CBRAM structure. The oxide consists of inhomogeneously distributed SiO_2 rings. The center of the SiO_2 rings are represented according to their number of ring-forming Si atoms as red, blue, and white spheres. The Ag filament (gray spheres) preferentially grows through regions with a sparser ring population, while it is prevented from entering dense regions with tight (5-fold) rings. (c) Zoom into the green rectangle in (b). It can be seen that the upper part (orange circle in (b)) of the filament has entered a pore (depicted by the red matrix) through an 8-fold SiO_2 ring (green). The bottom part (yellow circle in (b)) has grown through two 6-fold SiO_2 rings (blue) whereas tighter rings could not be passed.

A Dynamic Current Hysteresis Model for Thin-Film Transistors

Yu Li[1], Xiaoqing Huang[1], Congwei Liao[1], Runsheng Wang[2], Shengdong Zhang[1], Lining Zhang[1*], Ru Huang[2]
[1]School of ECE, Peking University, Shenzhen, China; [2]School of Integrated Circuits Peking University, Beijing, China
*Email: lnzhang@ieee.org

Abstract: This paper proposes a dynamic current hysteresis model for the Indium Gallium Zinc Oxide Thin Film Transistor (IGZO-TFT). Based on the Shockley-Read -Hall (SRH) theory, a kinetic equation that accurately describes the interface trap's capture/emission behavior is presented, which can incorporate the effect of interface trap density, trap energy level and scan rate dependency. Further, the kinetic equation is solved using a sub-circuit approach, combined with a validated TFT static current model, to achieve accurately simulating the current hysteresis of IGZO-TFT. This model has been validated with numerical TCAD simulations and has been shown to precisely reflect the effect of trap energy level, trap density and scan rate on the current hysteresis characteristics.

Keywords: Compact model, Current hysteresis, Indium Gallium Zinc Oxide Thin Film Transistor (IGZO-TFT)

Introduction: The progressive advancement of thin-film transistor technology has enabled the emergence of large-scale TFT circuits and integrated systems for various applications[1]-[3]. However, interface traps will be introduced during the processing of different types of TFT [4]-[7]. The trap induced degradation of the threshold voltage leads to the current hysteresis in the transfer characteristics and has a significant impact on the performance of the TFT device at the circuit level. For example, this hysteresis can be used to alter the brightness of the OLED[6]and make TFT array a potential contender for 3D stacked NVMs[3]. There are few reported works on modelling dynamic current hysteresis for circuit simulation based on trap dynamics. The design of large-scale systems places high demands on a current hysteresis model that captures the effects of the physical properties of the interface trap and the scan rate of the gate voltage.

Current hysteresis of IGZO-TFT: For acceptor-traps, "shallow" means that the trap energy level E_{tA} is close to the conduction band (i.e. $\Delta E_t = E_C - E_{tA}$ is relatively small). Fig.1(a) is the 3D structure of the IGZO-TFT device used in this paper. In numerical TCAD simulation, by introducing a certain amount of acceptor-traps ($\Delta E_t = 0.1\ eV, \mathrm{N}_{t,shallow} = 3\times10^{12} cm^{-2}$) at the IGZO/$SiO_2$ interface with device size of W = 20um and L = 20um, a clockwise hysteresis appears on the transfer characteristic curve (Fig1.(b)). The observed dependence of the hysteresis width on the scan rate is consistent with that reported in [6].

Dynamic Model Description: SRH theory [9] is often used to describe interface trap dynamics [10]-[11], this section discusses the implement of SRH theory in the trap dynamics simulation.

A. Trap Dynamics Based on SRH Theory:

Fig.2(a) illustrates the switching of a two-state trap state involving electrons in reservoir, where the transition rate (probability per unit time) $K_{12/21}$ is related to the capture/emission coefficients [12]. The trap occupancy probability $p_2(t)$ can be described as Eq (1), and the corresponding differential form of (1) can be rearranged as Eq (2). According to the SRH theory, the transition rates are shown in Eq (3), where $v_{th,n/p}$ denotes the carrier thermal velocity and $\sigma_{n/p}$ is the electron/hole capture cross section.

B. Sub-circuit Approach:

To facilitate the use of the sub-circuit approach to calculate the number of occupied acceptor-traps N_{tA}^-, Eq(2) can be reformulated as Eq(4), where N_t is the interface traps density with $N_{tA}^- = p_2(t)\cdot N_t$, and $\tau_{c/e}$ is the capture/emission time constant, and expressed as Equations (5)-(6).The sub-circuit shown as Fig.2(b) is built to solve the dynamic node voltage, then V_{th} is updated in the form of Eq (7).

Simulations and Results: Since the carrier capture cross section is related to the trap energy level, the time constant is found to depend on the gate voltage V_g and the defect energy level E_t from Eq (3) and (5). For better convergence and fit result of the dynamic model, different time constant expressions (shown as Eq (8)-(9)) are used in this work, while parameters are related to trap energy level and extracted respectively for shallow and deep energy level traps.

A. Threshold Voltage Bias Simulation:

Fig.3 shows the dependence of the time constant on the gate voltage for the shallow ($\Delta E_t = 0.1\ eV$) and the deep ($\Delta E_t = 0.6\ eV$) traps. The capture probability becomes larger when the trap energy level is bent below the Fermi energy level, creating a step-down trend in the capture time constant curve. During the simulation, the scan voltage is a triangular wave (Fig. 4(a)), while the influence of shallow and deep trap energy levels on the threshold voltage bias is given in Fig.4(b-c). It is worth noting that Fig.5(b) shows an overall upward shift in the threshold voltage bias curve for the deep traps, which is due to the earlier appearance of the step-down.

B. IGZO-TFT Static Current Model:

The credibility of the static current model is a prerequisite for the accuracy of the dynamic hysteresis model, however, there is no standard model of static current for IGZO-TFT at present[13]-[15]. In consideration of compatibility with trap kinetic behaviors, the open source threshold voltage based IGZO-TFT current model is used in this paper.[16] Fig.6 shows the agreement between the static current model and the current data from experiments and the TCAD reusults.

C. Current Hysteresis Simulation:

Combining the ΔV_{th} model and the static current model, the hysteresis curves at different scan rates (Fig.7) can be obtained. The hysteresis width of the shallow traps (Fig.7(a-c)) tends to increase and then decrease with decreasing scan rate. Fig.7(d-f) reveals that for the earlier appearance of the step-down causes the transfer characteristic curve to shift to the right as the scan rate decreases during forward scanning for deep traps. The scan rate dependence is quantified by using a fixed current method. For the current curves in Fig. 7, the difference of the gate voltage corresponding to the forward and backward sweep at Ids = 1nA is obtained as hysteresis width (Fig.8).

Conclusion: A dynamic current hysteresis model of IGZO-TFT is developed in this work. The trap dynamics described by the two-state trap model can be accurately solved by the sub-circuit method. Combined with the validated static current model, the dynamic model developed in this paper can well capture the effect of trap energy level, trap density on hysteresis width, while accurately reflecting the dependence of hysteresis width on scan rate.

Acknowledgement: This work is supported in part by the Natural Science Foundation of China (62074006), the Shenzhen Sci. and Tech. Project (20200827114656001) and (SGDX20201103095610029), and in part by the 111 Project (B18001).

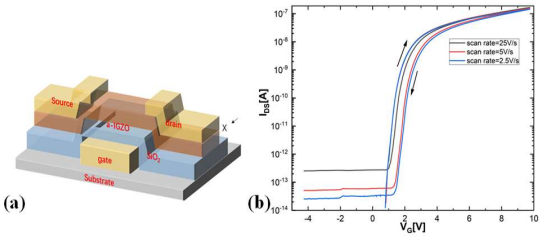


Fig.1. (a) Schematic of IGZO-TFT 3D structure. (b) Dependence of hysteresis width on scan rate.

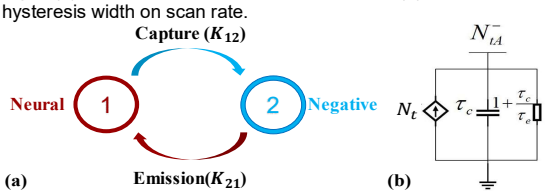


Fig.2. (a) Capture and emission events of Acceptor-trap. (b) Equivalent sub-circuit of the trap kinetic equation.

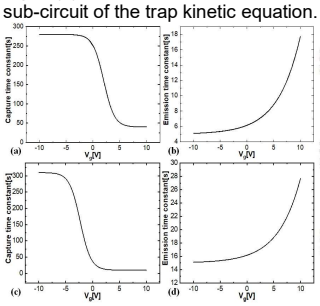


Fig.3. Capture and Emission time constant curves of shallow and deep energy level traps.

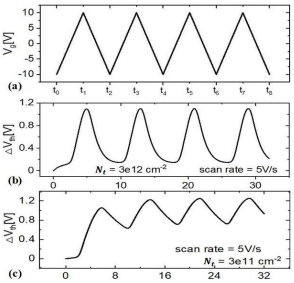


Fig.4. (a) The applied scan voltage. Threshold voltage bias dynamic responding of (b) shallow traps with $N_{t,shallow} = 3 \times 10^{12} cm^{-2}$. (c) deep traps with $N_{t,deep} = 3 \times 10^{11} cm^{-2}$.

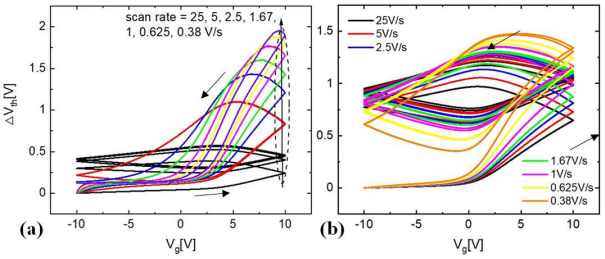


Fig.5. ΔV_{th} versus Vg for different scan rates of (a) shallow energy level traps ($\Delta E_t = 0.1eV$), (b) deep energy level traps ($\Delta E_t = 0.6eV$).

Equations:

$$p_2(t+\Delta t) = K_{12} \cdot \Delta t \cdot p_1(t) + (1 - K_{21} \cdot \Delta t)\, p_2(t) \quad (1)$$

$$\textit{Kinetic equation: } \frac{dp_2(t)}{dt} = K_{12} \cdot \left(1 - p_2(t)\right) + K_{21} \cdot p_2(t) \quad (2)$$

$$\textit{Transition rate: } K_{12} = v_{th,n} \cdot \sigma_n \cdot n;\ \ K_{21} = v_{th,p} \cdot \sigma_p \cdot p \quad (3)$$

$$\textit{Dynamics for Sub-circuit: } \frac{dN_{tA}^-(t)}{dt} = \frac{N_t - N_{tA}^-(t)}{\tau_c} - \frac{N_{tA}^-(t)}{\tau_e} \quad (4)$$

$$\textit{SRH time constant: } \tau_c = \frac{1}{K_{12}} = \tau_{c0} \cdot \exp\left(\frac{-V_g(t)}{n \cdot k \cdot T}\right) \quad (5)$$

$$\tau_e = \frac{1}{K_{21}} = \tau_{e0} \cdot \exp\left(\frac{V_g(t)}{n \cdot k \cdot T}\right) \quad (6)$$

$$V_{th}(t) = V_{th0} + \Delta V_{th}(t) = V_{th0} + \frac{q \cdot N_{tA}^-(t)}{C_{ox}} \quad (7)$$

$$\textit{Time constant model(with better convergence and fit): } \quad (8)$$

$$\tau_c = A_1 \cdot \tanh(B_1 \cdot V_G + C_1) + D_1$$

$$\tau_e = A_2 \cdot \exp(B_2 \cdot V_G + C_2) + D_2 \quad (9)$$

(parameters A-D are related to ΔE_t)

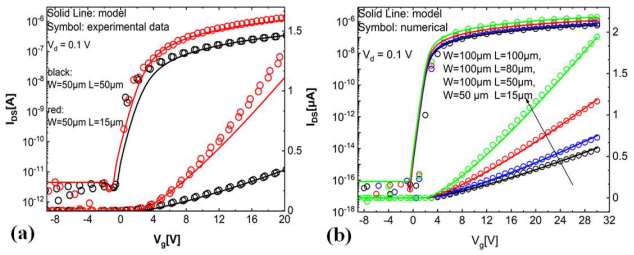


Fig.6. Agreement of the transfer characteristic between the static current models with (a) experimental data and (b) TCAD.

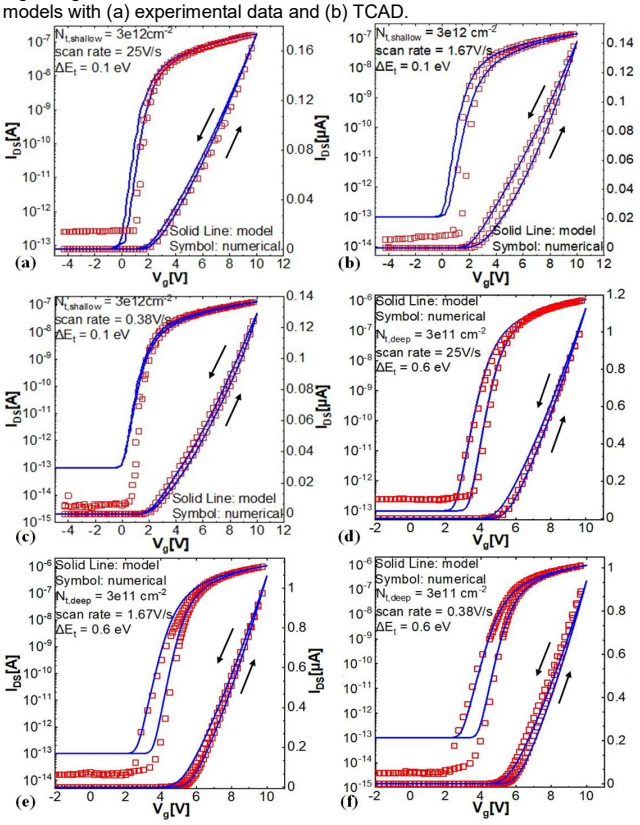


Fig.7. Model validation of the hysteresis characteristics of Id-Vg curves(W=20um, L=20um, $V_d = 1.1V$) for different scan rates for (a-c) shallow energy level traps ($\Delta E_t = 0.1eV$, $N_{t,shallow} = 3 \times 10^{12} cm^{-2}$ and (d-f) deep energy level traps ($\Delta E_t = 0.6eV$, $N_{t,deep} = 3 \times 10^{11} cm^{-2}$).

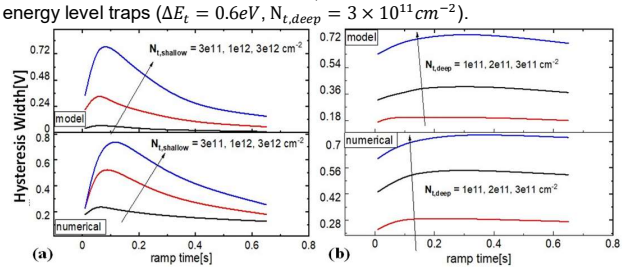


Fig.8. Dependence of hysteresis width on scan rate at different interface trap density for (a) shallow energy level traps, (b) deep energy level traps.

Reference: [1] R. Chaji et al., Thin Film Transistor Circuits and Systems. Cambridge Univ. Press, 2013. [2] S. Lee et al., Proc. IEEE, vol. 103, no. 4, pp. 644–664, Apr. 2015. [3] Z. Ye et al., *TED*, vol. 64, no. 2, pp. 438-446, Feb. 2017. [4] Sun Y et al., Org. Electron., 2015, 27: 192-196. [5] Lin H C et al., J. Appl. Phys, 2009, 105(5): 054502. [6] Y. Chen et al., *TED*, vol. 63, no. 4, pp. 1565-1571, April 2016. [7] Hung C H et al., MAT SCI SEMICON PROC, 2017, 67: 84-91. [8] Awawdeh K M et al., Org. Electron., 2013, 14(12): 3286-3296. [9] W. Shockley et al., Phys. Rev., vol. 87, no. 5, pp. 835-842, Sep. 1952. [10] T. Tsuchiya et al., Appl. Phys. Exp., vol. 4, no. 9, 2011. [11] P. Masson et al., APL., vol. 81, no. 18, pp. 3392-3394, 2002. [12] Grasser T et al., Microelectron Reliab, 2012, 52(1): 39-70. [13] Guo J et al., IEDM, 2020: 22.6. 1-22.6. 4. [14] Colalongo L et al., IEEE EDL, 2016, 37(4): 416-418. [15] Oodate Y et al., TED, 2015, 62(3): 862-868. [16] Shao L et al., IEEE Design & Test, 2019, 36(4): 6-14.

Analysis of 1/f and G-R Noise in Phosphorene FETs

Adhithan Pon, Avirup Dasgupta
DiRac Lab, Department of ECE, Indian Institute of Technology, Roorkee, Uttarakhand, India.
email: adhithan.ece@sric.iitr.ac.in; avirup@ece.iitr.ac.in;

***Abstract*—** This work deals with insights into low-frequency noise in phosphorene FET. In addition to the flicker noise component, which is often reported in the literature, we also look at generation - recombination (G-R) noise. We evaluate the dependence of noise on the number of layers for both armchair (AC) and zigzag (ZZ) orientations. We also extract the noise parameters for this device

I. Introduction

Two-dimensional Field Effect Transistors (2DFETs) are being actively investigated for future technology nodes[1]. For efficient use of these devices, understanding the noise behavior is of crucial importance[2]. Noise information also helps us understand the material better. For example, to understand the carrier dynamics in the 2D material systems, low-frequency noise spectroscopy is used as it provides information about trap energy levels and traps concentration values. This insight is highly useful for improving passivation and annealing processes. Of the various 2D materials being explored, phosphorene is currently receiving significant attention as it has a combination of useable bandgap and mobility values[3]. Many phosphorene-FET-based experiments have also been reported [4]. However, noise characterizations are quite limited, and a better understanding of noise dynamics is required. Especially for low-frequency noise, existing works usually neglect the impact of generation-recombination noise[5].

In this work, we develop an approach for low-frequency noise analysis in phosphorene FETs. This work uses first principle based simulations to calculate the material properties including all necessary physical effects while using TCAD for FET simulations. We analyze the noise behavior in Phosphorene FETs for multiple layers (N = 1 - 4) and both the orientations: armchair (AC) and zigzag (ZZ).

II. Method

A. DFT and Device simulation

We employ a semi-classical method to calculate the noise in Phosphorene FETs. First-principle based DFT simulations are used to attain the electrical properties of phosphorene layers (Fig. 2 & table 1). These are then used to perform the FET noise analysis. The detailed method/flow is presented in our previous works [6] (Fig. 1). The DFT method utilizes meta generalized gradient approximation (MGGA) and PBE functional to provide bandgaps for a different number of layers, which are calibrated with experimental data [7]. We leveraged 14 × 14 x 1 k-points sampling with a density mesh cut-off of 150 Rydberg for this calculation. Extensive work has been done to create a phosphorene material file for TCAD simulations using the result of the DFT simulations. The quantum effects are included using a 1D schrödinger equation with calibrated ladder parameters. Source and drain contacts are assumed to be perfect interfaces. Fig. 3 shows the calibration of our phosphorene FET against NEGF results [8].

B. Noise

To model the noise in the phosphorene FET, we use the impedance field method[9]. We consider diffusion, flicker, and generation-recombination (G-R). From the noise calibration (Fig. 4) the extracted Hooge parameter value is α_h = 5 x 10^{-4} [5] for bulk phosphorene whereas for silicon it is 2 x 10^{-3}. The flicker noise ($\propto 1/f^{\gamma}$) parameter $\gamma = 1.1$ and the trap density N_{trap} = 1 x 10^{12} cm^2. Layer and orientation-based noise parameters are also calculated while accounting for the thermal noise at high frequencies. Our noise spectrum shows a strong correlation with Hooge (mobility fluctuation)[10] and McWhorter (carrier fluctuation) models[11]. However, Van der Waals interaction and Source/Drain tunnelling are not considered.

III. Result and Discussion

The noise characteristics of phosphorene FET (N=1L, AC) is plotted as a function of frequency in Fig. 5 There are three main noise components dominant in three different frequency regions. For very low frequencies (10-1kHz) Generation – Recombination (G-R) noise is dominant due to the trapping and de-trapping of carriers. In the range of 1KHz – 100 MHz, the plot shows clear signs of flicker noise (1/f), which is due to carrier and mobility fluctuation in the phosphorene channel. For higher frequencies (greater than f_0 as marked in Fig. 5), the thermal white noise is dominant. Fig. 6, shows the extraction of the flicker noise ($1/f^{\gamma}$) parameter, γ. Fig. 7 shows the variation of γ and the factor $K = W \,.\, L \,.\, \alpha_h \,.\, I_{ds}^2$ (α_h is the Hooge parameter) with the number of layers as well as for different orientations (AC/ZZ). For ZZ, γ is roughly constant but there is a variation in K for different N, but for AC, γ varies while K is a constant. Fig. 8 shows the bias dependence of the noise power spectral density (S_{ID}) in phosphorene FETs. S_{ID} saturates at higher V_{gs} for all combinations (1L-4L & AC/ZZ). Also note that for the monolayer case the noise figure is constant for varying V_{ds} due to the very thin channel, whereas for N= 4L the noise figure increases with V_{ds} due to more inherent carrier fluctuation than the 1L. Fig. 9 shows that AC has slightly higher noise than ZZ orientation. This is because the higher conductivity for AC results in a higher thermal noise floor. This is true irrespective of the layer number. Fig. 10 summarizes the noise variation with the number of layers and different orientations. It also shows the Noise Figure (NF) trends in the inset.

IV. Conclusion

Through a detailed noise analysis for Phosphorene FETs, we have shown that S_{ID} increases with number of layers due to increased charge fluctuation. Also, ZZ has better noise performance than AC for all layer numbers. The Hooge model parameters are also extracted for the various combinations.

References

[1] M. Chhowalla, *et al*, *Nat. Rev.*, vol. 1, p. 16052, 2016. **[2]** A. A. Balandin, *Nat. Nanotechnol.*, vol. 8, no. 8, pp. 549–555, 2013. **[3]** L. Li *et al.*, *Nat. Nanotechnol.*, vol. 9, no. 5, pp. 372–377, 2014. **[4]** N. Antonatos *et al.*, *ACS Appl. Mater. Interfaces*, vol. 12, no. 6, pp. 7381–7391, Feb. 2020. **[5]** W. Liu *et al.*, *Nanophotonics*, vol. 9, no. 7, pp. 2053–2062, 2020. **[6]** A. Pon, *et al.*, *IET Circuits, Devices Syst.*, vol. 14, no. 8, pp. 1167–1172, 2020, **[7]** J. Qiao, *et al.*,*Nat. Commun.*, vol. 5, pp. 1–7, 2014. **[8]** X. Cao *et al.*, *IEEE TED*, vol. 62, no. 2, pp. 659–665, 2015. **[9]** T.Synopsys,SDEVICE Manual. [10] F. N. Hooge., *Phys. Lett. A*, vol. 29, no. 3, pp. 139–140, Apr. 1969. **[11]** E. Burstein, *et al.*, Philadelphia (Pa.) : University of Pennsylvania press, 1957.

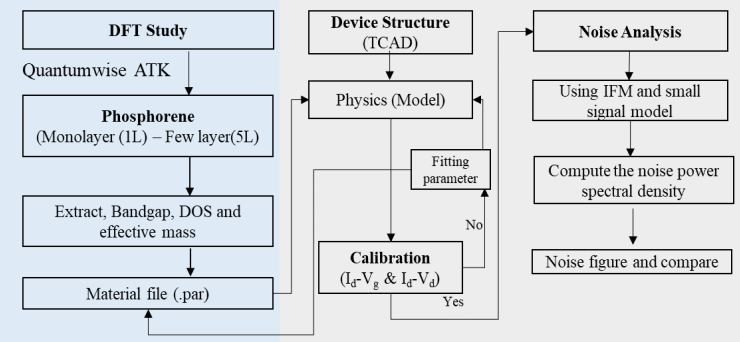


Fig. 1. Simulation flow of phosphorene FET drain characteristics and noise performance analysis

Table:1 Material parameters used for simulation

Parameters	Values
Affinity (eV)	3.9 – 4.15
Ladder parameters A,B,C	4.1;0.7;23.05
KVM,α_t,α_b-AC (ZZ)	15(20) ; 0.1(2); 20(0.3);
Workfunction (eV)	4.25
Hooge	$1x10^{-5}$ - $5 x10^{-5}$
Trap con.	$1 x 10^{12}$ /cm^2
EOT (nm)	1
1L- Nce_DOS (cm^{-2})	$8.42 x 10^{16}$
5L- Nce_DOS(cm^{-2})	$9.77 x 10^{16}$
Ae	24.7
α_h	$5 x 10^{-4}$
$\tau 1$, $\tau 0$	1, $1 x 10^{-6}$

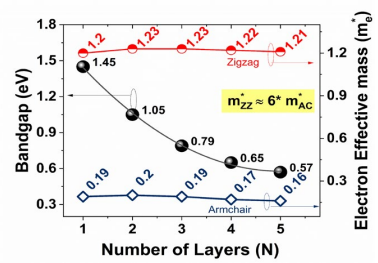


Fig. 2. The phosphorene bandgap and effective mass values are plotted as a function of the number of layers. AC orientation has less effective mass than ZZ. The bandgap is reduced as the number of layers increases due to reduction in confinement.

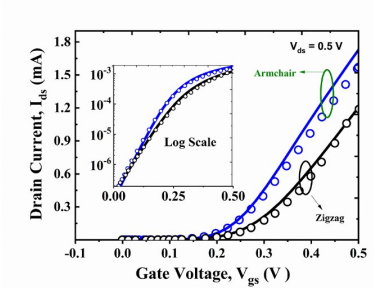


Fig. 3. The calibration of BP-FET against NEGF results [8]. The calibrated parameters are listed in table 1. Source/drain doped with n-type (4×10^{18}cm−3) and their extension is 14 nm. The contacts are assumed ohmic and trap free.

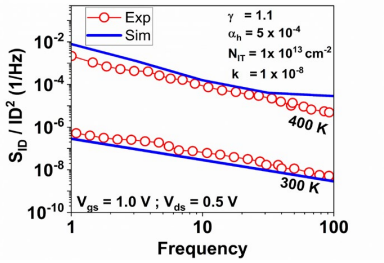


Fig:4 Noise spectral density (S_{ID}) for bulk phosphorene as a function of frequency and calibrated with [5]. Calibrated noise parameters values are given in the insets. This indicates perfect scaling of $1/f^{\gamma}$ as expected in conventional noise. This trend is affected by the number of layers considered.

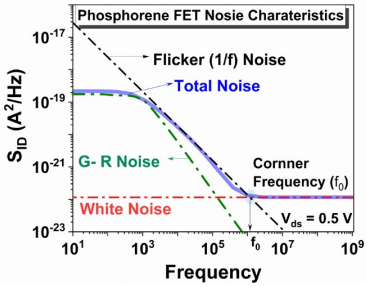


Fig:5 Noise characteristics of Phosphorene FET (1L-AC). It is found that at low frequencies S_{ID} is high due to a combination of flicker (1/f) and G-R noise. The corner frequency (f_o) is marked, indicating the dominance of the thermal noise floor.

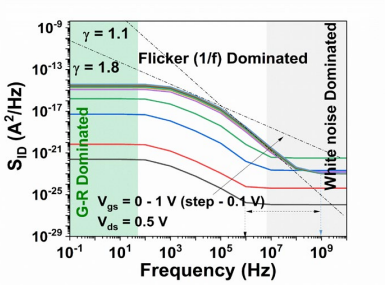


Fig:6 S_{ID} of phosphorene device (N= 1L AC) as a function of frequency for a range of gate biases V_{gs} = 0.0 – 1 V and V_{ds} = 0.5 V. S_{ID} is saturated after V_{gs} ~ 0.4 V due to the onset of strong inversion. γ is also a function of bias.

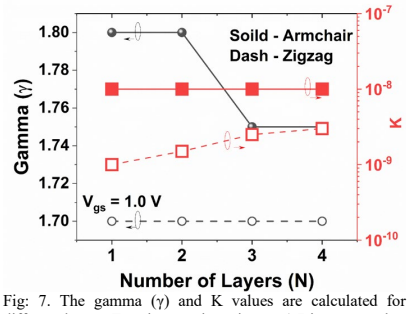


Fig: 7. The gamma (γ) and K values are calculated for different layers. For zigzag orientation γ= 1.7 is constant but K varies with number of layers. For armchair orientation, on the other hand, K is constant but γ varies with number of layers. $S_{ID} = K/f.N$ where $K = W.L.\alpha_h.I_{ds}^2$

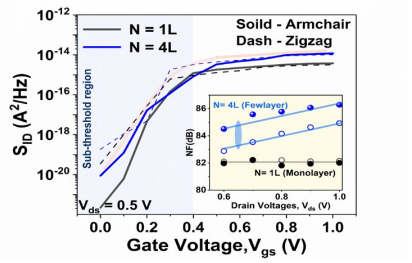


Fig:8 Noise spectrum plotted as a function of gate & drain bias. It is clearly shown up to the subthreshold region (> 0.4) S_{ID} increases exponentially due to more G-R and diffusion noise. However, drain bias dependence varies with the number of layers. For N=1L the noise values saturate perfectly due to the very thin nature, whereas for N= 4L it keeps increasing due to more inherent carrier fluctuation than the 1L.

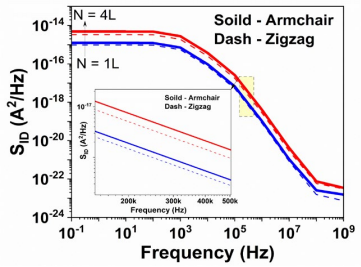


Fig:9 Noise spectrum for different layers. The thermal noise in the AC direction is high due to its high conductivity compared to ZZ. The favorable transport direction AC (high μ and less m_e^*) creates more thermal noise resulting in the overall noise to be higher for all frequencies compared to ZZ.

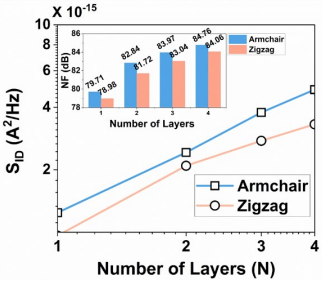


Fig:10 S_{ID} for phosphorene with different layers. In particularly ZZ has good noise performance in the wide frequency range. The inset figure shows the noise figure in the dB with different layers. The layer dependency of noise grows as the number of layers increases; the reason for this is, that a multilayer has more volume noise than a monolayer.

Microstructural Impact on Electromigration Reliability of Gold Interconnects

H. Ceric, R. L. de Orio, and S. Selberherr

Institute for Microelectronics, TU Wien, Gußhausstraße 27-29, 1040 Wien, Austria
Email: {ceric|orio|selberherr}@iue.tuwien.ac.at

Simulating the influence of microstructure on the electromigration (EM) reliability of a metallic interconnect is a challenging task for two main reasons: firstly because of the complexity of the vacancy behavior within and near grain boundaries; and secondly because of the large number of grain boundaries. In some applications the nano-scale of the interconnect width causes the emergence of larger grains extending through the whole interconnect width. The presence of large grains significantly simplifies modeling of the microstructural impact. However, in the case of gold interconnects used for GaAs [1, 2] technology we have to deal with a significant number of small grains (∼10) through the whole depth of the interconnect. Gold is the metal of choice for interconnects implemented on GaAs, because it forms a very low resistance ohmic contact, has a high melting temperature, and a low resistivity [3]. In the recent study on EM reliability of interconnects made of gold by Hau-Riege and Yau [4] the relevant segments of metallization contained several thousands of grains. The layout of the used test structure is shown in Fig. 1. Grain boundaries act not only as sites of vacancy recombination, but also, depending on the angle they close to the current flow direction, they either enhance (low angle) or diminish (high angle) vacancy transport. Utilizing both the multi-physics simulation tool COMSOL [5] as well as the microstructure and mesh generator NEPER [6], the physical effects of microstructure can be included in reliability models to a high level of detail. In COMSOL, the mechanical and the material-transport model can be assigned to each grain or to groups of grains fully taking in account the local material properties, like diffusivity, elasticity parameters, and crystal orientations (cf. Fig. 2). Each of the entities consisting of a single or of a multitude of grains is separately meshed (cf. Fig. 3). In this work a novel approach for the numerically efficient simulation of an interconnect containing thousands of grains is introduced. A number of grains are grouped into larger grain compounds and at the boundaries of these domains a detailed grain boundary model is applied [7]. The boundaries of the large grain compounds are chosen to comprise characteristic segments of the interconnect geometry (e.g., vias). Alternatively, these segments can also be divided into several grain compounds. The size statistics of grain compounds are set according to the experimentally obtained statistics of grain size distribution. Inside the compounds the effective diffusivity (D_{eff}) is calculated as a weighted sum of contributions from the grain bulk (D_{bulk}) and from the grain boundaries (D_{gb}).

$$D_{\text{eff}} = D_{\text{bulk}} + \delta \frac{\sum_i S_i}{\sum_j V_j} D_{\text{gb}} \tag{1}$$

δ is the grain boundary width and S_i and V_i are the grain boundary surfaces and volumes, respectively. The summation is performed along all the grains contained in a compound. The microstructure modeling is integrated in the usual modeling framework for simulation of EM reliability [8, 9], which comprises the solution of the vacancy balance equation together with the Laplace equation, the heat-transport equation, and the mechanical equations. The simulation provides as output, for an arbitrary three-dimensional multilayer interconnect geometry and for given initial and boundary conditions: the temperature distribution, the current density distribution, the tensorial mechanical stress field, and the point defect (vacancy) distribution. These values enable the determination of the interconnect lifetime. The capability of the novel simulation approach is verified through comparison with the previously published [4] experimental results. The failure time statistics of gold interconnects are obtained through testing in a MIRA EM module [2] for different interconnect dimensions. The experimentally observed EM lifetime distributions for all experiments were monomodal and lognormally distributed. In Fig. 4 the simulated lifetime distributions for four different via

sizes are presented. As one can see, all distributions are lognormal and monomodal like in experiments [4]. Additionally, the experimentally observed dependence of the mean failure time and the associated standard deviation of the failure times on the interconnect geometry is also well reproduced by our simulations.

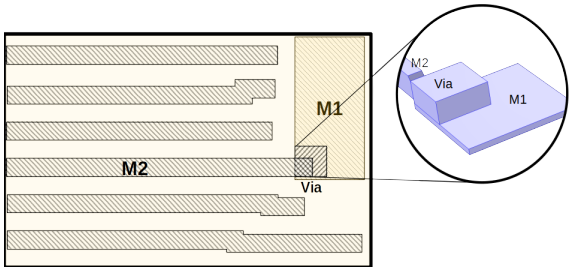


Figure 1: Layout of the line-end of the investigated test structures [4].

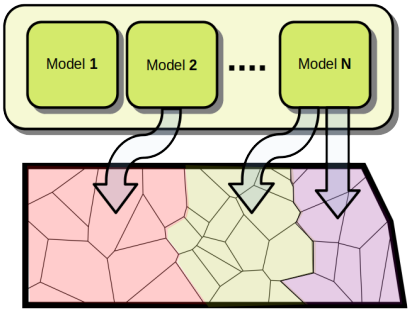


Figure 2: Schematic picture of the overall modeling approach. From the model library of N configured models, available in the simulation tool, two models are chosen and assigned to the three different grain compounds.

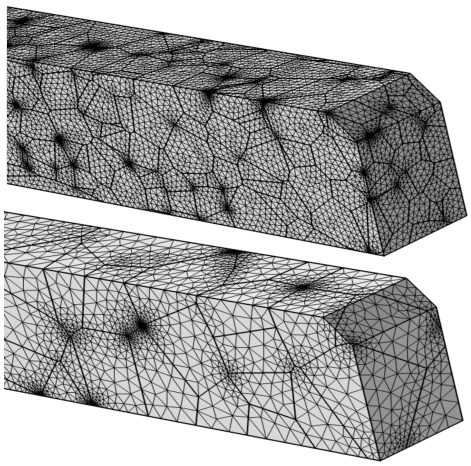


Figure 3: Example of meshing of an interconnect segment for two different grain size distributions.

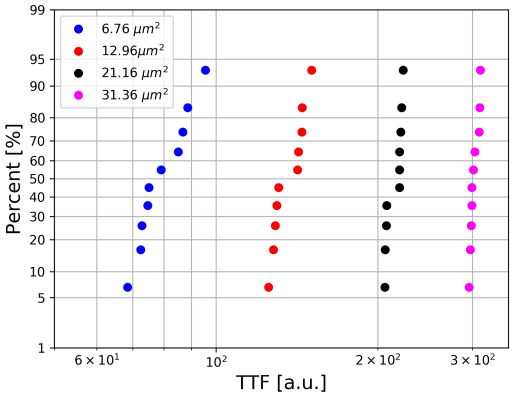


Figure 4: Lifetime distribution for different via sizes.

REFERENCES

[1] N. A. Kulchitsky, A. V. Naumov, and V. V. Startsev, *Mod. Electron. Mater.*, vol. 6, p. 77, 2020.
[2] S. Kilgore, Ph.D. dissertation, Arizona State University, 2013.
[3] S. Kasap and P. Capper, *Springer Handbook of Electronic and Photonic Materials.* Springer, 2017.
[4] C. Hau-Riege and Y. Yau, in *Proc. Intl. Symposium on the Physical and Failure Analysis of Integrated Circuits*, 2021, pp. 1–5.
[5] COMSOL Multiphysics, Version 5.6, 2021.
[6] R. Quey, P. Dawson, and F. Barbe, *Comput. Methods. Appl. Mech. Eng.*, vol. 200, no. 17-20, pp. 1729–1745, 2011.
[7] H. Ceric, H. Zahedmanesh, and K. Croes, *Microelectron. Reliab.*, vol. 100-101, p. 113362, 2019.
[8] M. E. Sarychev and Y. V. Zhitnikov, *J. Appl. Phys.*, vol. 86, no. 6, pp. 3068–3075, 1999.
[9] H. Ceric and S. Selberherr, *Mater. Sci. Eng. R Rep.*, vol. 71, pp. 53–86, 2011.

Reliability of TCAD Study for HfO_2-doped Negative Capacitance FinFET with Different Material Specific Dopants

Rajeewa Kumar Jaisawal, Sunil Rathore, P.N. Kondekar, and Navjeet Bagga*
VLSI Design and Nano-scale Computational Lab, Electronics and Communication Engineering Department, PDPM-IIITDM Jabalpur, India- 482005 (*Email: navjeet@iiitdmj.ac.in)

Abstract: Attaining the ferroelectric (FE) polarization in a thin HfO_2 layer using a specific dopant is a widely adopted way to realize Negative Capacitance (NC) FET. In a general TCAD simulation study of NC-based devices, the NC property of the FE layer is strongly dependent on the values of Landau parameters $(\alpha, \beta, \gamma, \rho, g)$, which are unique for specific dopants and FE thickness. In this paper, for the first time, we investigated the reliability of TCAD simulations with which NC FinFET is simulated for specific dopants-based FE-HfO_2 layer. The possible dopants used in the thin-HfO_2 layer are Al, Gd, La, Si, Sr, Y, and Zr. Each dopant has different $(\alpha, \beta, \gamma, \rho, g)$ and thus offers a different NC regime of operation, i.e., S-curve. α, and β are the dominant parameters if we consider the uniform polarization under quasi-static analysis. Further, the change in ambient temperature alters the value of α, resulting in changes in the NC-state. Hence, for the reliable TCAD-based NC study, the precise selection of Landau parameters and dopants is needed for optimized performances.

Introduction: In sub-22nm node, FinFETs-based CMOS devices have gained popularity due to enhanced gate electrostatic controllability [1]. However, the physical mechanism of the carrier transport imposes a limit of 60mV/decade on the subthreshold slope (SS). Thus, to realize a steeper-slope device, the idea of NC was reported and implemented using FE-layer in the gate stack [2]-[3]. In downscaled devices, the realization of the conventional perovskite materials is not feasible due to their larger thicknesses. Thus, the HfO_2 doped layer is implemented to realize a thin film FE layer for compatible process integration [4]. To the best of our knowledge, in the available literature, to create a thin film doped-HfO_2 FE layer, seven different dopants are used: D_1: Aluminium (Al) [P_r=5, E_C=1.3]; D_2: Gadolinium (Gd) [P_r=20, E_C=1.75]; D_3: Lanthanum (La) [P_r=45, E_C=1.2]; D_4: Silicon (Si) [P_r=10, E_C=1.0]; D_5: Strontium (Sr) [P_r=23, E_C=2.0]; D_6: Yttrium (Y) [P_r=24, E_C=1.2]; and D_7: Zirconium (Zr) [P_r=18, E_C=1.0]. The P_r and E_C are in $\mu C/cm^2$ and MV/cm, respectively. Thus, in TCAD simulation of NC devices, the dopant-dependent Landau parameters are taken. However, a thorough investigation of the reliability and optimization of these parameters, i.e., the choice and significance of the dopants for the HfO_2 layer to convert into the FE layer, is not yet explored. *The key contribution of this work is*: (*i*) investigation of the impact of Landau parameters $(\alpha, \beta, \gamma, \rho, g)$ for uniform FE layer under quasi-static operation; (*ii*) impact of temperature on α and its significance on the NC-regime, i.e., S-curve; (*iii*) the impact of temperature on the transconductance and drivability over different dopant-dependent FE layer; (*iv*) capacitance matching and its variation with varying temperature. Thus, a proper investigation is needed to select the dopant of the HfO_2-layer for optimal NC operation.

TCAD Setup for NC-FinFET: A 14nm node industry-standard nFinFET is employed to realize the baseline NC-FinFET (Fig.1) by placing the Zr-doped HfO_2 layer in the gate stack in our simulation study using *Sentaurus* TCAD [5]. The TCAD setup includes a conventional drift-diffusion model for carrier transport. The mobility, saturation velocity, SRH recombination, high field saturation, quantum correction models, etc., have been included and adequately tuned to get a good match between the TCAD and experimental [6] I_{DS}-V_{GS} curves (Fig.2a). The device parameters used in the simulation are mentioned in Table-I. Further, to realize the NC effect in FinFET, the FEPolarization model is incorporated and calibrated with an experimental MFIM (metal-ferroelectric-insulator-metal) capacitor [7] (Fig.2b). The extracted α, β, γ parameters are used to get a good fit of the S-curve (Fig.2c). The realized NC-FinFET show improved performances (Fig.2d-e), i.e., I_{ON} & SS, due to the internal voltage amplification of the FE-layer.

Results and Discussion: In TCAD simulations, the NC effect can be included using the FEpolarization model with appropriate Landau parameters, which are dependent on the dopant of the HfO_2-based FE layer. We have considered the default value of ρ=2.25×10^4 Ω-cm, γ=0, and g=10^{-4}cm^3/F, as mentioned in the TCAD manual under quasi-static analysis for single domain FE-layer. Therefore, the vital parameters are α and β, which significantly influence the NC state. For different dopants, we plotted the S-curves, showing the NC regime (Fig. 3a). D_2 offers a larger span of the NC region; however, D_3 shows higher NC. Whereas the I_{OFF} degrades in D_3 due to the lower value of β, in turn, higher SS (Fig.3c,e). In contrast, the enhanced I_{ON} is achieved in D_4 and D_6, showing a higher dependency of I_{ON} on α (Fig.3b,d & Table-II). The impact of temperature (T) is severe on α, as it is a linear function of T (Fig.4). The increase in T reduces the coefficient α_{0x} (Fig.4a), resulting in a change in α_{xT} (x: dopant). Therefore, the NC regime gets shrinks (Fig.4b). The I_{ON} thus reduces with higher T (Fig.4c). An increase in T also causes mobility degradation and the bandgap narrowing; in turn, I_{OFF} increases (Fig.4d). The impact of temperature and Landau parameters are essential for analog design metrics like transconductance (g_m), channel capacitance (C_{gg}), etc. As we increase the T, the g_m of the doped HfO_2-based NC FinFET will decrease (Fig.5a-e). The impact of β on threshold voltage (V_{th}) is more pronounced (Fig.5d). The capacitance matching with varying Landau parameters and temperature is investigated in Fig.6. The capacitance matching occurs at the subthreshold region for the dopants with higher P_r, i.e., lower α and β. However, the capacitance matching in the saturation region dominates for the dopants having lower values of P_r, i.e., higher α and β. Further, the temperature influences the capacitance matching in NC FinFET, resulting in the modulation of channel capacitances (Fig. 6a-e).

Conclusions: For a reliable TCAD simulation of the Negative Capacitance (NC) devices, the selection of the appropriate Landau parameters $(\alpha, \beta, \gamma, \rho, g)$ is crucial as it frames the span and slope of the S-curve, which states the NC operation. In scaled NC devices, the HfO_2-layer is doped with the appropriate dopants to achieve ferroelectricity. The type of the dopants alters the Landau parameters and thus the reliability of the simulated NC devices. For quasi-static operation under consideration of a single domain FE layer, only the α and β parameters play a significant role. The I_{ON} and I_{OFF} are functionally dependent on α and β, respectively. Further, the temperature shows a linear trend with α and strongly influences α_0 (i.e., temperature coefficient). Thus, an increase in temperature deteriorates the device's performance. Hence, a thorough investigation of $(\alpha, \beta, \gamma, \rho, g)$ is essential to simulate an NC device in TCAD.

Acknowledgment: N. Bagga greatly acknowledge the support obtained from PDPM-IIITDMJ (project no. IIITDMJ/ODRSPC/2022/88)

References: [1] S.Y. Wu et al., IEDM 2016. [2] S. Salahuddin et. al., Nano lett. 2008. [3] R. K. Jaisawal et.al, MEJ 2021. [4] M.G. Kozodaev et. al., Appl. Phys. Lett. 2017. [5] *Sentaurus* TCAD Manual. [6] C. -H. Lin et al., IEDM 2014. [7] M. Hoffmann et. al., IEDM 2018.

Reliability of TCAD Study for HfO2-doped Negative Capacitance FinFET with Different Material Specific Dopants

Motivation: a) Material specific parameter dependent Negative Capacitance (NC) analysis using TCAD.
b) Analyzing the change in doping dependent Landau parameter with temperature for optimized NC operation.

Obj. 1: Variance of L-K parameters for different material specific dopants
Obj. 2: Impact of temperature on L-K parameters for different material specific dopants
Obj. 3: impact of temperature on the transconductance and drivability
Obj. 4: Capacitance matching analysis for different material specific dopants and their undertaken operating conditions.

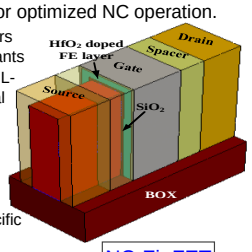


Fig.1. Schematic of an NC-FinFET realized by placing a HfO_2-doped ferroelectric (FE) layer in the gate stack.

Calibration of TCAD Setup

TABLE-I

Parameter	Value
Gate Length	20 nm
EOT	0.9 nm
Spacer length (L_{sp})	8 nm
Fin height	26 nm
Fin width	6.5 nm

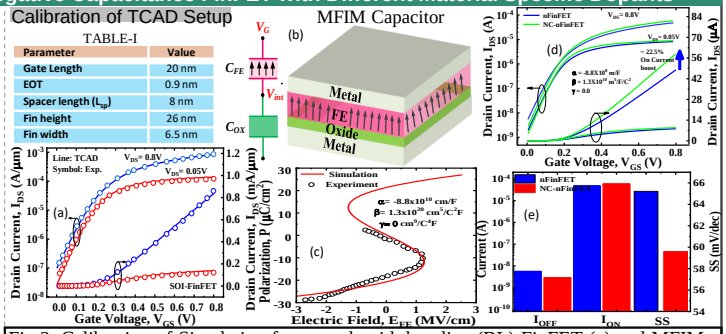


Fig.2. Calibration of Simulation framework with baseline (BL) FinFET (a) and MFIM capacitor (b-c). Realized NC-FinFET (Fig. 1) offers higher I_{ON}, lower I_{OFF} and lower subthreshold slope (SS) [d-e]. Parameter used in simulation are mentioned in Table-I.

Obj.1: L-K parameters for Material Specific Dopants in HfO_2-based FE layer

$$E_{FE} = 2\alpha_x P + 4\beta_x P^3 + 6\gamma_x P^5 + \rho_x \frac{\partial P}{\partial t} - 2g_x \Delta P$$

$$\alpha_x = -\alpha_x' \times 10^{10} cm/F \quad \beta_x = \beta_x' \times 10^{20} cm^5 / F / C^2$$

$$\gamma_x = 0,\ g_x = 1\times10^{-4} cm^3 / F \text{ (for uniform domains)},\ \rho_x = 2.25\times10^4 \Omega - cm \text{ (for quasi-static analysis)}$$

P: Polarization
E_{FE}: Field across FE layer
$\alpha_x, \beta_x, \gamma_x, \rho_x, g_x$: L-K parameters which governs the NC phenomenon in TCAD

The possible dopants for HfO_2-based FE layer are: *i*) Aluminum (Al); *ii*) Gadolinium (Gd); *iii*) Lanthanum (La); *iv*) Silicon (Si); *v*) Strontium (Sr); *vi*) Yttrium (Y); *vii*) Zirconium (Zr)

TABLE:II

Dopants	α_x'	β_x'
Al (D1)	10.55	2.06
Gd (D2)	11.36	1.42
La (D3)	3.46	0.8
Si (D4)	12.99	6.49
Sr (D5)	11.29	1.06
Y (D6)	12.5	3.28
Zr (D7)	8.6	1.3

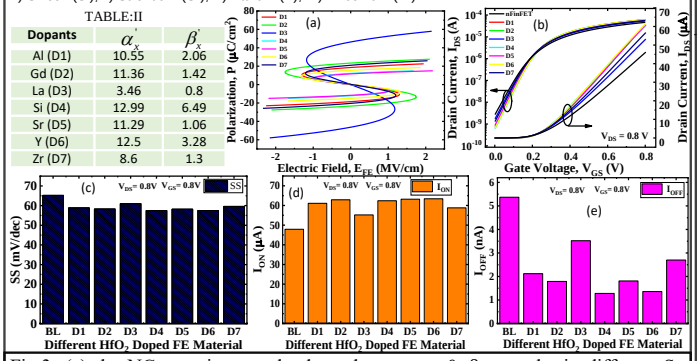


Fig.3. (a) the NC-state is strongly dependent on α_x & β_x, results in different S-curves for different α_x' and β_x' values, as mentioned in Table-II; (b) I_D-V_{GS} characteristics for different material specific dopants; (c) SS for different dopants; (e-f) I_{ON} (I_{OFF}) dominantly governs by $\alpha_x'(\beta_x')$.

Obj.2: Impact of Temperature on L-K parameters

$$\alpha_x = \alpha_{0x}(T_R - T_C)$$

$$\Rightarrow \alpha_{xT} = \alpha_{0x}(T - T_C)$$

T_C: Curie Temp. (<650 K); above which FE property may lost
T_R : Room Temp.; T : Temp. (250K to 450K)
α_{0x} : Positive Temp. Coefficient

β_x, γ_x are weaker function of temperature

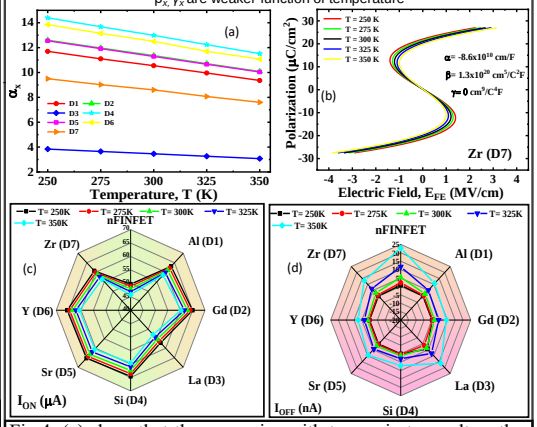


Fig.4. (a) show that the α_{xT} varies with temp., in turn, alters the FE polarization; (b) S-curve shows the influence of temp. on NC-state of operation; (c) influence of temp on I_{ON} and I_{OFF} (d).

Obj.3: Impact of Temperature on g_m

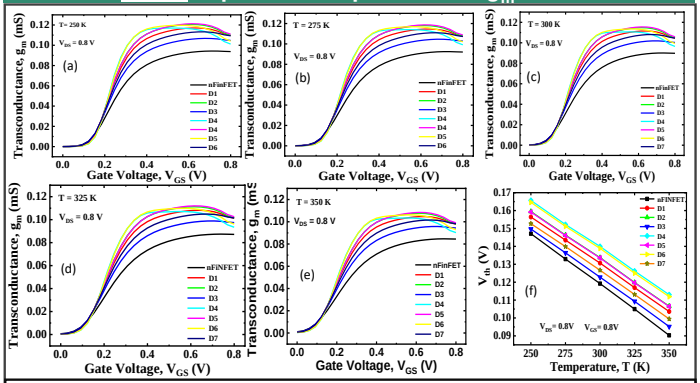


Fig.5. shows the impact of increasing the ambient temperature from 250K to 350K on transconductance (g_m) (a-e), which is indeed a key parameter in analog circuit design. Increasing the temperature decrease the maximum g_m value due to the deterioration of dominant α Landau parameter. However, the temperature-induced g_m and threshold voltage (V_{th}) deterioration (f) are mitigated by internal voltage amplification provided by the dopant-dependent HfO_2-based FE layer.

Obj. 4: C-V Characteristics for Different Dopants

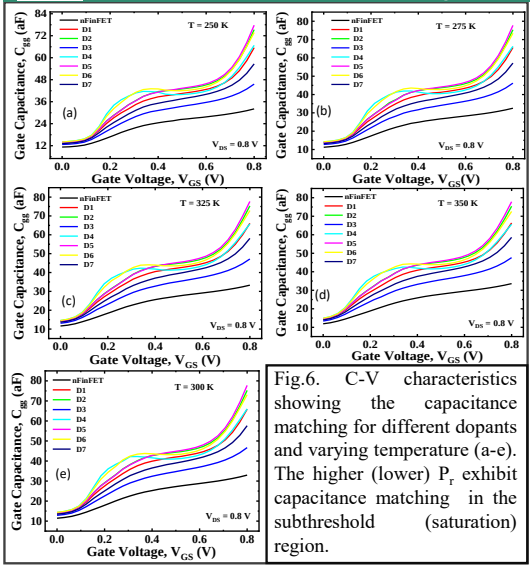


Fig.6. C-V characteristics showing the capacitance matching for different dopants and varying temperature (a-e). The higher (lower) P_r exhibit capacitance matching in the subthreshold (saturation) region.

A proposal of quantum computing algorithm to solve Poisson equation for nanoscale devices under Neumann boundary condition

Shingo Matsuo and Satofumi Souma†
Department of Electrical and Electronic Engineering, Kobe University, Kobe 657-8501, Japan
†email: ssouma@harbor.kobe-u.ac.jp

***Abstract*—We present an implementation study of gate-type quantum computing algorithms for the purpose of semiconductor device simulations. As one of the representative quantum algorithms we consider the use of HHL (Harrow-Hassidim-Lloyd) algorithm to solve the Poisson equation in semiconductor nanowire p-n junction, especially under the Neumann boundary condition that the electric field is zero at the electrode boundaries. Our proposed model of the quantum gate to implement the Neumann boundary condition has been found to successfully reproduce the solution obtained by conventional method.**

I. Introduction

Recent progress of quantum computing technology, especially the dramatic progress of the quantum computing environment via the cloud such as in IBM Q, have been stimulating various studies on the application of intermediate-scale quantum computers, so-called NISQ (Noisy Intermediate-Scale Quantum) devices. Such research on the specific application examples of quantum computers ranges from material science simulations to social science simulations such as finance. In any of these cases, it is important to consider how quantum computers can be used in solving equations that have been required in individual fields.

In semiconductor device simulation as well, it is important to consider the possibility of utilizing quantum computing algorithms in the long term in the future. One of such examples is the use of HHL (Harrow-Hassidim-Lloyd) algorithm [2], which is one of the applications of the quantum phase estimation algorithm, in the calculation of the potential distribution based on Poisson equation [3], [4]. With such motivation we study how the HHL algorithm can be applied to solve the Poisson's equation for semiconductor nanowire structure, especially under the Neumann boundary condition that the electric field is zero at the electrode boundaries.

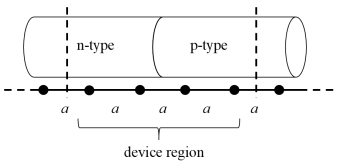


Fig. 1. Scematic illustration of the semiconductor nanowire p-n junction system, where the left and the right regions are doped into n and p types, respectively. Finite differentiated grids are also illustrated.

978-1-6654-4200-8/21/$31.00 ©2021 IEEE

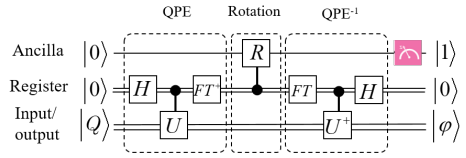


Fig. 2. Quantum gate circuit of HHL algorithm to solve matrix Poisson equation $A|\varphi\rangle = |Q\rangle$, where A is capacitance matrix and is encoded in the unitary operation U. The quantum circuit is composed of Hadamard gate (H) quantum Fourier transformation (FT), rotation (R), and unitary operation (U). If the ancilla (top most) qubit is measured to be $|1\rangle$, the solution $|\varphi\rangle$ of the linear equation is encoded in the bottom set of the qubits.

II. Proposed method and results

In this study we assume the semiconductor nanowire p-n junction system shown in Fig. 1, where the left and the right regions are doped into p and n type, respectively. Assuming that the electrostatic potential and the charge density is constant within the cross-sectional area S and thus depend only on the longitudinal position x, the Poisson equation is written as $\varepsilon d^2\varphi(x)/dx^2 = -\rho(x)$, where ε is the dielecric constant, $\varphi(x)$ is the electrostatic potential, and $\rho(x)$ is the charge density. By introducing the finite difference approximation, the discretized Poisson equation is derived as $-C\varphi_{i+1}+2C\varphi_i-C\varphi_{i-1} = Q_i$, where i stands for the position grid along the x direction, $C \equiv \varepsilon S/a$ is the capacitance with a being the grid spacing, and $Q_i \equiv \rho(x_i)\,Sa$ is the charge at the ith grid. In order to concentrate our focus on how can we apply the HHL algorithm for the device simulation purpose, in this study, we restrict our attention to the simplified model, where the central device region is described by four grid points spanned by $i = 1 \sim 4$, and semi-infinite leads are attached at both ends of the central region. By imposing the Neumann boundary condition that the electric field is zero at both ends of the central device region, so that $\varphi_1 = \varphi_0$ and $\varphi_5 = \varphi_4$, we obtain the matrix equation

$$\begin{pmatrix} C & -C & & \\ -C & 2C & -C & \\ & -C & 2C & -C \\ & & -C & C \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \end{pmatrix} = \begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \end{pmatrix}. \quad (1)$$

This linear equation has the form of $A|\varphi\rangle = |Q\rangle$, where the vectors $|\varphi\rangle$ and $|Q\rangle$ are expressed by quantum bits (qubits),

and the capacitance matrix A is expressed by quantum gates in the HHL algorithm shown in Fig. 2 as we will explain next.

In the HHL algorithm shown in Fig. 2 the unitary operation (U) is defined by the time evolution due to the Hamiltonian matrix A, so that $U = e^{iAt}$. In order to implement this unitary evolution by quantum gate circuit, it is necessary to decompose the matrix A as follows.

$$\begin{aligned} A &= \begin{pmatrix} \varepsilon_1 & \gamma & 0 & 0 \\ \gamma & \varepsilon_2 & \gamma & 0 \\ 0 & \gamma & \varepsilon_2 & \gamma \\ 0 & 0 & \gamma & \varepsilon_1 \end{pmatrix} \\ &= \begin{pmatrix} \varepsilon_1 & 0 & 0 & 0 \\ 0 & \varepsilon_2 & 0 & 0 \\ 0 & 0 & \varepsilon_2 & 0 \\ 0 & 0 & 0 & \varepsilon_1 \end{pmatrix} \\ &+ \begin{pmatrix} 0 & \gamma & 0 & 0 \\ \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma \\ 0 & 0 & \gamma & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma & 0 \\ 0 & \gamma & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &= A_0 + A_1 + A_2, \end{aligned} \tag{2}$$

where $\varepsilon_1 = C$, $\varepsilon_1 = 2C$, and $\gamma = -C$. Here we note that A_0 and A_1 are operators within a single qubit, while A_2 is essentially the operator connecting two qubits (inter qubit operator). The corresponding unitary evolution operator is then calculated by applying the Trotter expansion as

$$U = e^{iAt} = e^{i(A_0+A_1+A_2)t} \simeq \left(e^{iA_0t/n}e^{iA_1t/n}e^{iA_2t/n}\right)^n, \tag{3}$$

where larger value of n gives more accurate results. This unitary evolution operator can be implemented by the quantum gate circuit in Fig. 3. Here we note that since the matrix A_0 is not the multiplication of the identity matrix due to the Neumann boundary condition, we propose to use the quantum gate model depicted in the $e^{iA_0t/n}$ part in Fig. 3. As for the quantum gate circuits for the quantum Fourier transformation (FT in Fig. 2) and the controlled rotation (R in Fig. 2) we employ the standard gate configurations [3].

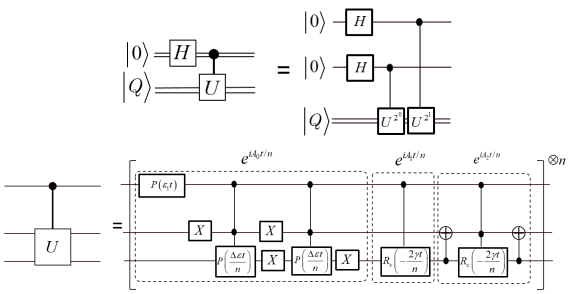


Fig. 3. Detail of the quantum circuit corresponding to the controlled unitary U part in Fig. 1. P abd R_x are the phase and x rotation gates, respectively

Here it should be noted that the eigenvalues of the matrix A are $E = 0$, $2-\sqrt{2}$, 2, and $2+\sqrt{2}$, where the energy difference between adjacent eigenvalues is $\sqrt{2}$ except for the lowest one

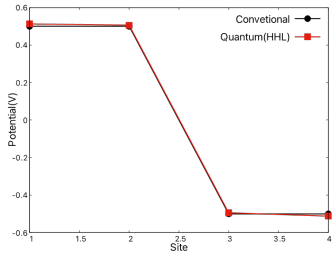


Fig. 4. Comparison of the electrostatic potentials obtained by HHL algorithm and conventional (LU decomposition) method. Here we assume that the charges of $+0.26e$ and $-0.26e$ at the 2nd and the 3rd sites, respectively. Our parameter values correspond to the case that the charge density is $\rho = 10^{26}e$ m^{-3}, grid spacing is $a = 2.6$ nm, cross-sectional area is $S = 1$ nm^2, and the material is Si. The value of n in Trotter expansion has been chosen as 128, and the time constant is $t = 2\pi \times 0.25/\sqrt{2}$.

(we note that the lowest eigenvalue is zero and thus should be excluded to calculate the solution $|\varphi\rangle$). We can then take the advantage of this fact if the diagonal elements of the matrix A is subtracted by the constant $\Delta E \equiv 2 - 2\sqrt{2}$, because then the 2nd, 3rd, and 4th eigenvalues $E' = E - \Delta E$ are exactly described by binary numbers 0.01, 0,10, and 0,11, respectively, where 0.01 corresponds to the decimal expression 0.25 and this is scaled from the eigenvalue $E' = \sqrt{2}$ by the factor $0.25/\sqrt{2}$. The subtracted value ΔE can be re-incorporated at the controlled rotation part R to obtain the correct solution $|\varphi\rangle$. We implemented the above scheme using the quantum programming language Qiskit, and obtained the solution $|\varphi\rangle$ very close to that obtained by the conventional method as shown in Fig. 4 (see caption for detailed parameters).

III. Conclusion

We have present an implementation study of gate-type quantum computing algorithms for the purpose of semiconductor device simulations. As one of the representative quantum algorithms we considered the use of HHL (Harrow-Hassidim-Lloyd) algorithm to solve the Poisson equation in semiconductor nanowire p-n junction, especially under the Neumann boundary condition that the electric field is zero at the electrode boundaries. Our proposed model of the quantum gate circuit to implement the Neumann boundary condition has been found to successfully reproduce the solution obtained by conventional method. Although the above explanation has been restricted to the case of four sites model, the proposed basic idea to obtain the accurate solution can be directly applicable to 2^n sites model in general by appropriately designing the non-uniform grid to obtain the equally distanced eigenvalues.

References

[1] S. Souma and M. Ogawa, IEICE Electronics Express, 17.20190739 (2020).
[2] A. W. Harrow, A. Hassidim, and S. Lloyd, Phys. Rev. Lett. **103**, 150502 (2009).
[3] S. Wang, Z. Wang, W. Li, L. Fan, Z. Wei, and Y. Gu, Quantum Inf. Process. **19**, 170 (2020).

[4] H. Morrell and H. Y. Wong, SISPAD2021 (2021).

Approximate H-Transformation for Numerical Stabilization of a Deterministic Boltzmann Transport Equation Solver Based on a Spherical Harmonics Expansion

Suhyeong Cha and Sung-Min Hong
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology,
123 Cheomdan-gwagiro (Oryong-dong), Buk-gu, Gwangju, 61005, Republic of Korea
E-mail: smhong@gist.ac.kr

Abstract

There are two widely adopted stabilization schemes for a deterministic Boltzmann transport equation solver based on a spherical harmonics expansion. One is the maximum entropy dissipation scheme and the other is the H-transformation. In this work, we propose an alternative numerical stabilization method, which is based on the approximate H-transformation. In the proposed scheme, a new energy variable approximately follows the total energy. An additional term is generated out of the free-streaming operator and it should be implemented properly. When the kinetic energy is fixed, the distribution function at that energy can be directly accessible at any time instance. The proposed scheme is implemented in our in-house deterministic Boltzmann transport equation solver. The numerical simulation results demonstrate that the proposed stabilization scheme works properly without any numerical difficulties.

Introduction

Deterministic Boltzmann transport equation solvers for the three-dimensional electron gas [1]-[4] have been actively studied. Recently, the transient simulation results using an implicit time marching technique have been reported in [5]. In [5], the maximum entropy dissipation scheme [3] was adopted. The H-transformation, another numerical stabilization scheme widely adopted, is not very convenient to be used in the transient simulation, because the distribution functions at previous time instances must be interpolated for the present potential profile. Therefore, the transient simulation capability that is compatible with the H-transformation has not been reported yet.

In this study, in order to overcome the difficulties originated from the stabilization scheme, an alternative stabilization scheme, which is based on the approximate H-transformation, is proposed.

Approximate H-Transformation

It is assumed that the energy variable is uniformly discretized with a spacing, ΔE. For the sake of notational simplicity, the dependence on the position variable is omitted. Let us consider a case where an implicit time marching scheme is adopted to discretize the time derivative term. At a given kinetic energy, ε, the time derivative of Zf (a product of the density-of-states, Z, and the distribution function, f) at a time instance, t_i, can be written as

$$\left.\frac{\partial[Zf]}{\partial t}\right|_{t=t_i} \approx Z(\varepsilon)\sum_{j=0}^{N-1} a_{i,j} f\left(\varepsilon, t_{i-j}\right), \qquad (1)$$

where N is the order of the time-marching scheme, $a_{i,j}$ is a coefficient connecting t_i and t_{i-j}. For example, the simplest backward Euler scheme ($N = 2$) has $a_{i,0} = -a_{i,1} = \frac{1}{t_i - t_{i-1}}$.

The conventional H-transformation uses the total energy, H, which is defined as

$$H = \varepsilon + V = \varepsilon + (-q\phi + E), \qquad (2)$$

where V is the band minimum energy, q is the absolute elementary charge, ϕ is the electrostatic potential, and E is the energy difference between the band minimum and the local reference energy. Variable transformation from ε to H greatly simplifies the free-streaming operator of the Boltzmann transport equation. However, when the transient simulation is involved, the terms related with the time derivation introduce some difficulties. Since $V(t_i) - V\left(t_{i-j}\right)$ does not vanish in general, $f\left(\varepsilon = H - V(t_i), t_{i-j}\right)$ is not directly available. An interpolation procedure must be introduced. From the present authors' experience, with a finite resolution of numerical discretization, the interpolation procedure deteriorates the numerical stability.

An alternative approach based on the approximate H-transformation is proposed. Instead of the time-varying band minimum energy, its approximation, $\tilde{V}$, is used to contruct a new energy variable, h,

$$h = \varepsilon + \tilde{V}. \qquad (3)$$

Under this new transformation, an additional term appears out of the free-streaming operator. Fortunately, since this additional term is proportional to $V - \tilde{V}$, which is small for a reasonable $\tilde{V}$, a suitable discretization technique can be developed.

The major progress made in this work is to propose a condition for $\tilde{V}$, with which the interpolation procedure in the time derivation can be completely eliminated. We propose a condition of

$$\tilde{V}(t_i) = k(t_i)\Delta E, \qquad (4)$$

where $k(t_i)$ is an integer function and ΔE is a constant energy spacing. Moreover, $k(t_i)$ is unambiguously determined to minimize the absolute value of $V - \tilde{V}$. With the above condition, we readily have

$$\tilde{V}(t_i) - \tilde{V}\left(t_{i-j}\right) = \left[k(t_i) - k\left(t_{i-j}\right)\right]\Delta E. \qquad (5)$$

Again, the time derivation requires $f\left(h - \tilde{V}(t_i), t_{i-j}\right)$. Since the energy shift is just an integer-multiple of ΔE, the required distribution function can be readily accessible without any interpolation procedure. Therefore, with the approximate H-transformation, the transient simulation becomes possible while enjoying the superior numerical properties of the conventional H-transformation.

Numerical Results

The proposed method has been newly implemented into our in-house deterministic Boltzmann transport equation solver. The physical models are the same with those in [5]. Three different stabilization schemes (the maximum entropy dissipation scheme, the conventional H-transformation, and the approximate H-transformation) are available.

Figure 1 shows the structure under simulation. Starting from a 1200-nm-long N^+NN^+ resistor, a scaled structure can be generated by increasing a scaling factor. As shown in Fig. 2, the DC IV characteristics of all simulated structures show excellent agreement between two methods. It is much expected because the approximate H-transformation is a modified version of the conventional H-transformation.

The distribution function at equilibrium is drawn in Fig. 3. Two stabilization schemes are compared. As expected, the conventional H-transformation follows the Boltzmann distribution, which is not dependent on the position variable. On the other hand, with the approximate H-transformation, the distribution function varies over the position variable, because the energy variable slightly differs from the total energy.

When a high DC voltage is applied to the anode terminal, the stabilization scheme based on the kinetic energy suffers from negative distribution functions. In Fig. 4, the color map of the distribution is shown. Empty areas at high energies represent discretized points with negative distribution functions. For the same case, as shown in Fig. 4(b), the distribution function calculated with the approximate H-transformation is non-negative everywhere.

Conclusions

In conclusion, a novel stabilization scheme has been successfully implemented into the deterministic Boltzmann transport equation solver. The distribution functions at past time instances can be directly accessible without any interpolation procedure and the stability of the conventional H-transformation can be shared. The proposed method can be readily applied to the transient simulation.

References

[1] A. Gnudi et al, SSE, vol. 36, pp. 575-581, 1993.
[2] K. A. Hennacy et al., SSE, vol. 1485-1494, 1995.
[3] C. Jungemann et al., JAP, vol. 100, p. 024502, 2006.
[4] S.-M. Hong and C. Jungemann, JCE, vol. 8, pp. 225-241, 2008.
[5] S.-M. Hong and J.-H. Jang, JEDS, vol. 6, pp. 156-163, 2018.

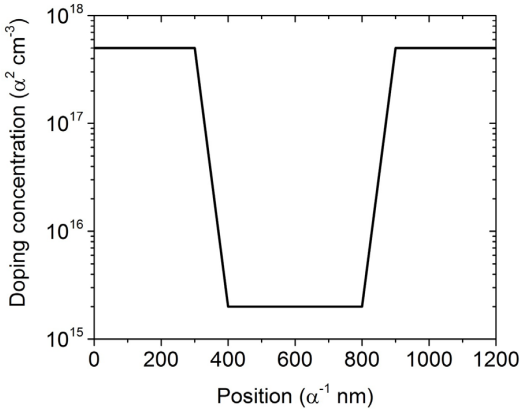


Fig. 1. Doping profile of a N^+NN^+ resistor simulated in this work. By increasing a scaling factor (α), various structures can be generated. In this work, the scaling factor varies from 1 (1200-nm-long structure) to 10 (120-nm-long structure).

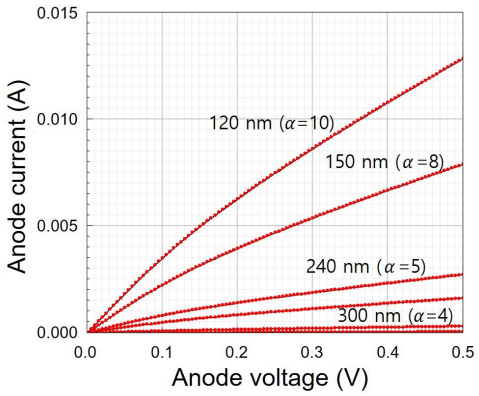


Fig. 2. DC IV characteristics of different structures with various scaling factors. The calculation results with the conventional H-transformation (Black lines) and the approximate H-transformation (Red dots) are almost identical.

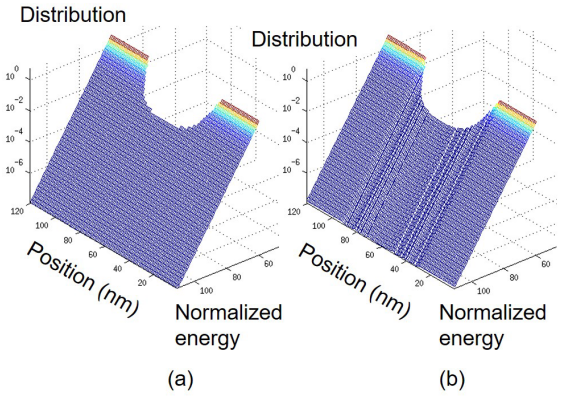


Fig. 3. Distribution function at equilibrium calculated with (a) the conventional H-transformation or (b) the approximate H-transformation. The total length is 120 nm. At a given position variable, each distribution function decays exponentially as the energy variable increases. The additional term out of the free-streaming operator plays an important role in the approximate H-transformation.

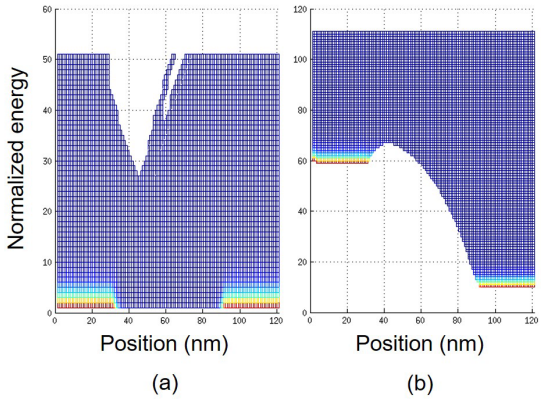


Fig. 4. Color map of the distribution function calculated with (a) the maximum entropy dissipation scheme or (b) the approximated H-transformation. The total length is 120 nm and the applied anode voltage is 0.5 V. Negative distribution functions are clearly observed in (a). Of course, an integral of the distribution function along the energy axis is positive and areas with negative distribution functions can be reduced by adopting a finer grid. On the other hand, in the case of the proposed methods, the distribution function is non-negative everywhere, even with the same grid.

Coupling a phase field model with an electro-thermal solver to simulate PCM intermediate resistance states for neuromorphic computing

O. Cueto[a], A. Trabelsi[a], C.Cagli[a], M.C Cyrille[a]
[a]Univ. Grenoble Alpes,CEA, LETI, DCOS, LSM F-38000 Grenoble; E-mail: olga.cueto@cea.fr;

I. Introduction

Phase-Change Memories (PCM) are today considered the most mature among novel non-volatile memory technologies. PCM rely on the reversible and rapid phase transition between the amorphous and crystalline phases of chalcogenide materials such as $Ge_2Sb_2Te_5$ (GST225). The ability to alter the conductance levels in a controllable way makes PCM devices particularly well-suited for synaptic realizations in neuromorphic computing [1] [2] [3]. The PCM devices are currently being investigated as building blocks for neuromorphic circuits [4]. A key attribute that enables this application is the progressive crystallization of the phase-change material and subsequent increase in device conductance by the successive application of appropriate electrical pulses. We studied by simulation the progressive crystallization of GST225 PCM. The simulations are realized with a dedicated tool that relies on the coupling of the Phase Field Method (PFM) with an electro-thermal solver [5]. The devices studied are GST225 state-of-the-art PCM devices. The simulations presented in this work were compared to electrical results and show a qualitative agreement for the simulated conductance evolution during the first recrystallization pulses. Our model allows understanding the mechanims of recrystallization during set pulses and consequently can help controlling the intermediate states of resistance of PCM devices.

II. A solver for PCM coupling the Phase-Field method with an electro-thermal model

The finite element electro-thermal solver coupled with the PFM was already presented [5] and we just give here a reminder of its main features. The PFM relies on an order parameter η representing the local crystallinity of PCM material (η=1 in the crystalline phase and η=0 in amorphous phase). Time evolution of η is governed by the Allen-Cahn equation (1) and corresponds to the reduction of the total energy of the system composed of free-energy of bulk phases and energy of interfaces between phases [6].

$$\frac{\partial \eta}{\partial t} = -L_\eta \left(\frac{\partial f(\eta, T)}{\partial \eta} - \kappa \nabla^2 \eta \right) \tag{1}$$

$f(\eta, T)$ is a local free-energy density for which we use the expression proposed by [7]. Phase change mechanisms that occur during the PCM operations generally involve concurrent nucleation and growth. Our nucleation model already presented [8] relies on the Classical Nucleation Theory (CNT) [9]. Our approach is to introduce nucleation model as a complementary entity from the PFM equation and to have two algorithms which alternate, one for nucleation and one for growth and coarsening corresponding to the advance of equation (1). During the nucleation steps, critical nuclei are introduced into individual cells randomly.

An ohmic model approach is used to simulate the electrical behavior of the PCM cell. The electro-thermal solver relies on the coupled system of partial differential equations formed by the current conservation equation and the heat transfer equation.

$$\nabla \cdot (-\sigma \nabla V) = 0 \tag{2}$$

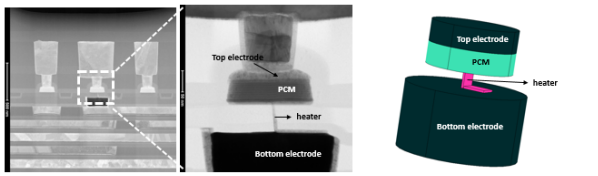
The heat equation in the PCM material includes the energy exchange due to the latent heat of melting.

$$\rho C_p \frac{\partial T}{\partial t} + \nabla \cdot (-k_{th} \nabla T) = \sigma (\nabla V)^2 + L \frac{dh}{d\eta} \frac{\partial \eta}{\partial t} \tag{3}$$

where σ, ρ, C_p, k_{th} and L stand for the PCM electrical conductivity, density, heat capacity, thermal conductivity and latent heat of melting and h is a smooth interpolation function. For PCM, σ and k_{th} depend on η and T. More precisely, σ is implemented using a state variable with four possible values (crystal, liquid, amorphous with high or low electrical resistivity) and takes into account electronic switching. The amorphous phase is electrically in the off-state if the local electric field is lower than a threshold electric field (E_{th}) and switches to the on-state if the local electric field becomes higher than E_{th} [10]. The system of previous equations is discretized using the Finite-Element method with the Partial Derivative Equations interface of COMSOL multiphysics ® [11].

III. Simulation and Measurement results and discussion

We performed experiments on a GST-based PCM array. The memory array consists of 16kbit single bank built on 28nm CMOS technology front end on 300mm wafer. The selector devices are MOS transistors with thin oxide. The PCM resistor is a GST225 PCM device integrated into the LETI Memory Advanced 300mm Demonstrator (Fig. 1). The electrical measurements were carried out with an Agilent B1530 [12]. First a reset pulse is applied, which leads to a full reset state ($R > 10^6\Omega$). Next a series of 20ns\10μs\20ns (rise\width\fall times) were delivered to the cell. The selector gate is used to control the current. Two regimes were studied: a high current regime ($I_h = 120\mu A$) and a low current ($I_l = 25\mu A$). After each pulse, the resistance of different 10 devices is averaged and plotted in Fig. 2. As can be seen the conductance (G) increases firstly linearly (pulses $< 20 - 30$) and then saturates when



a- TEM images (global and Zoom) b- schematic view of the PCM structure

Fig. 1: Illustration of the PCM device.

additional pulses are applied. This can be partially explained by the drift of the amorphous phase that happens during the period between successive set pulses.

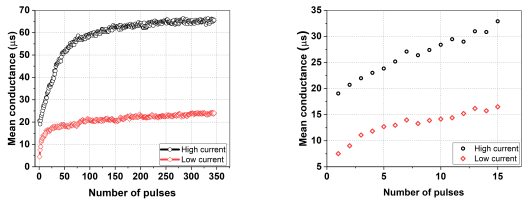


Fig. 2: G versus number of pulses - left: for 350 pulses - right: for the first fifteen pulses.

Starting from a device with an amorphous dome obtained by the reset pulse simulation, we have qualitatively reproduced the progressive increase of conductance associated to low and high current set pulses. The mechanisms of recrystallization can be nucleation or growth dominated depending on the set pulse characteristics. Our simulations clearly indicate a nucleation dominated mechanism in the conditions of our set pulses. As

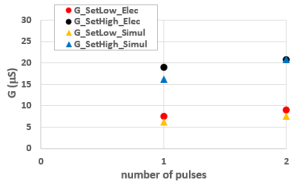
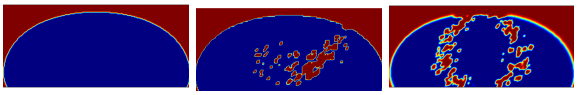


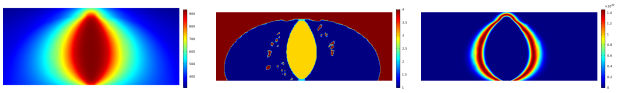
Fig. 3: comparison of G by simul. versus elect. charact. for the two first pulses.

plotted in Fig. 3, the increase of conductance associated to each set pulse is reproduced by the simulation for the two first pulses. For more pulses, the simulation model overestimates the increment of conductance because currently it does not integrate the drift phenomenon which is supposed to counterbalance the increase of conductance [12].
In the low current regime, neither fusion nor re-amorphization occurs during the set pulse and increase of conductance is due to nucleation and growth within the amorphous dome as illustrated by Fig. 4 b. We hypothesize that the conductivity saturation can still be ascribed to the counter effect of drift even though no-reamorphization occurs. For the set pulses with higher current, the scenario of recrystallization is different because the current is high enough to trigger the melting of the central part of the active domain leading to a nucleation rate maximal in the periphery of the melted domain as visible in Fig. 5. As indicated by the simulation (Fig. 4 c), the high current regime corresponds to a re-amorphization of the GST mushroom core, while the surrounding GST slowly nucleates, leading to the increase of conductance. It can be shown that in this scenario, the conductivity saturation can be fully explained by the counter effect of resistance drift of the amorphous, which tends to lower the conductance after each pulse [12].



(a) after Reset (b) Reset then Set Low (c) Reset then Set High

Fig. 4: Two crystallization regimes for Set Low and Set High pulses: crystalline phase(purple red), amorphous phase(blue).



(a) Temperature (b) State (c) Nucleation Rate

Fig. 5: Pulse Set High t=400ns b) the yellow domain corresponds to the melted PCM c) nucleation rate is maximal on the periphery of the melted domain.

IV. CONCLUSION

The progressive crystallization that occurs during the successive application of set pulses was studied by electrical characterization and simulation. Insight on the real mechanims of crystallization is obtained by coupling Joule heating, diffusion of heat and a phase change model. We reproduce and explain the two different regimes of recrystallization corresponding to set low and set high pulses evidenced by electrical characterization. Thanks to this model the hypothesis that drift counterbalances the increase of conductance during the progressive application of set pulses is confirmed. This work paves the way to a better control of the set pulses to be used to generate predefined conductance levels in the PCM for neuromorphic computing.

V. ACKNOWLEDGEMENT

A.Trabelsi is supported by the CEA NUMERICS program, which has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 800945.

REFERENCES

[1] C. D. Wright et al, Adv. Funct. Mater. 2013, 23, 2248-2254
[2] S.R. Nandakumar et al, Journ.of Applied Physics 124,152123,2018
[3] A. Sebastian et al 2019 J. Phys. D: Appl. Phys. 52 443002
[4] D. Yiğit et al, proceedings of IEEE conf. ISCAS 2021
[5] O. Cueto et al, proceedings of IEEE conf. SISPAD 2015
[6] S.M. Allen and J. W. Cahn, Acta Metallurg. vol. 27, no.6, Jun. 1979
[7] Y.Kwon et al, IEEE Electron Device Letters Vol.34, no.3, March 2013
[8] A.Gliere et al, proceedings of IEEE conf. SISPAD 2011
[9] K.F Kelton et al, J. Chem. Phys., vol. 79, no. 12, 1983
[10] O.Cueto et al, proceedings of IEEE conf. SISPAD 2012
[11] Comsol mutiphysics 5.5 (www.comsol.fr)
[12] A. Trabelsi, submitted to IEEE conf. ESSDERC 2022

Discontinuous Galerkin Concept for Quantum-Liouville Type Equations

1st Valmir Ganiu
Chair for High Frequency Technique
TU Dortmund
Dortmund, Germany
valmir.ganiu@tu-dortmund.de

2nd Dirk Schulz
Chair for High Frequency Technique
TU Dortmund
Dortmund, Germany
dirk2.schulz@tu-dortmund.de

***Abstract*—A time-dependent discontinuous Galerkin method for the numerical solution of the Liouville-von-Neumann equation for the analysis of quantum transport in nanoelectronics and nanophotonics systems in center mass coordinates is presented. With this methodology, a further increase in computation efficiency is achieved compared to conventional methods, particularly when considering large-scale problems.**

***Index Terms*—computational nanotechnology, numerical methods, quantum transport, von-Neumann equation, Quantum Liouville type equation.**

I. Introduction

Recently, a method for the analysis of the carrier transport within quantum devices was proposed that starts from the Liouville-von Neumann equation in center-mass coordinates and applies a Finite Volume (FV) method for the spatial approximation [1]. After an expansion of the density matrix according to predefined eigenfunctions, an equation is obtained, which corresponds to the conventional Quantum Liouville Equation (QLE). With this method, also called Quantum Liouville type equation (QLTE), suitable boundary conditions are incorporated by the introduction of a complex potential. The latter approach is essential to obtain physically meaningful results [1]. Alternatively, instead of choosing the FV technique, an implementation applying Finite Element (FE) techniques would be conceivable. For time-dependent problems, it must be noted that FV schemes or even conceivable FE techniques require the solution of large equation systems. Thus, such algorithms are exceedingly computationally expensive. Alternatively, discontinuous Galerkin (DG) methods have proven themselves in fluid dynamics as for example, which bypass the solution of equation systems and instead allow a computation via parallelizable matrix-vector multiplications. Due to the mathematical relationship of the initial master equations, e.g. Navier-Stokes equations and QLTE, the DG method is also a suitable technique for the approximation of QLTEs. This algorithm is presented, validated and the computational efficiency is evaluated.

II. Discontinuous Galerkin Approach

For demonstration purposes, the DG approach is presented based on a one-dimensional problem. The framework is the Liouville-von-Neumann equation (LVNE) utilizing a spatially constant effective mass Hamiltonian and Hartree-Fock potential [2], [3]. The LVNE is then transformed into center mass coordinates χ and ξ. The von-Neumann in center mass coordinates reads as

$$\frac{\partial}{\partial t}\boldsymbol{\rho}(\chi,\xi,t) = i\frac{\hbar}{m}\frac{\partial}{\partial\chi}\frac{\partial}{\partial\xi}\boldsymbol{\rho}(\chi,\xi,t) - i\frac{q}{\hbar}\boldsymbol{B}(\chi,\xi,t)\rho(\chi,\xi,t). \tag{1}$$

Here, $\boldsymbol{\rho}$ denotes the statistical density matrix and the term $\boldsymbol{B}(\chi,\xi,t)$ contains the Hartree-Fock potential. Next, the computational domain Ω_ξ within the interval $[-\frac{L_\xi}{2},+\frac{L_\xi}{2}]$ is subdivided into cells. Eq. (1) is approximated with respect to the coordinate ξ utilizing an equidistant grid with N_ξ cell points. The resulting relation is conceptually characterized by matrices $\boldsymbol{A}$ and $\boldsymbol{G}$, both showing a dimension of N_ξ x N_ξ:

$$\frac{\partial}{\partial t}\boldsymbol{\rho}(\chi,t) = \boldsymbol{A}\frac{\partial}{\partial\chi}\boldsymbol{\rho}(\chi,t) - \boldsymbol{G}(\chi,t)\boldsymbol{\rho}(\chi,t) \tag{2}$$

Ultimately, a basis transformation is required to allow a distinction between forward and backward propagation of waves. Thus, basis vectors $\boldsymbol{\Phi_n}$ together with their corresponding expansion coefficients c_n are introduced. The basis vectors are orthonormal, such that $\boldsymbol{\Phi}_n^\dagger\cdot\boldsymbol{\Phi}_m = \delta_{n,m}$. Accordingly, the transformation

$$\boldsymbol{c}(\chi,t) = \boldsymbol{\Phi}^\dagger\cdot\boldsymbol{\rho}(\chi,t) \tag{3}$$

is applied with $\boldsymbol{\Phi}$ containing the basis functions $\boldsymbol{\Phi}_n$. Exploiting this transformation, the QLTE

$$\frac{\partial}{\partial t}\boldsymbol{c}(\chi,t) = \boldsymbol{\Phi}^\dagger\boldsymbol{A}\boldsymbol{\Phi}\frac{\partial}{\partial\chi}\boldsymbol{c}(\chi,t) - \boldsymbol{\Phi}^\dagger\boldsymbol{G}(\chi,t)\boldsymbol{\Phi}\cdot\boldsymbol{c}(\chi,t) \tag{4}$$

is obtained, with $\boldsymbol{\Lambda} = \boldsymbol{\Phi}^\dagger\boldsymbol{A}\boldsymbol{\Phi}$ being a diagonal matrix containing the eigenvalues of $\boldsymbol{A}$. To solve (4), an approximation in χ-direction is needed. For this purpose the DG concept is applied, whereas the computational domain Ω_χ resides in the interval $[0, L_\chi]$. To account for a discrete solution for each element k in the χ-direction, test functions $l_i(\chi)$ must be introduced. Consequently, (4) is multiplied with each of the defined test functions. The resulting relations are integrated with respect to χ for each element k defined in the domain $[D^k] = [\chi_l^k, \chi_r^k]$, where χ_l^k and χ_r^k are the coordinates on the left and right edge of each element, respectively. As for standard FE approaches, a stiffness matrix $\boldsymbol{S}^k$ and a mass matrix $\boldsymbol{M}^k$ are introduced. In contrast to these standard approaches a

single-valued numerical flux $f_j^{k,*}(\chi,t)$ is introduced to ensure the stability of the DG scheme and to establish the coupling between elements. For this purpose, (4) is integrated twice utilizing Green's theorem to arrive at the so called local strong formulation

$$\boldsymbol{M}^k \partial_t \underline{c}_j^k - \boldsymbol{S}^k \underline{f}_j^k + \boldsymbol{M}^k \sum_{m=1}^{N_\xi} \mathrm{diag}(\underline{G}_{jm}^k)\underline{c}_m^k = \oint_{\partial D^k} (f_j^k(\chi,t) - f_j^{k,*}(\chi,t)\hat{n}\underline{l}(\chi). \quad (5)$$

Contrary to hydrodynamics, within quantum transport problems, the reactive part dependent on the matrix $\mathrm{diag}\underline{G}_{jm}^k$ corresponding to the drift term $\boldsymbol{B}$ must be considered when evaluating (5). For demonstration purposes, an upwind flux is chosen to account for the forward and backward waves. Its values are defined as

$$\oint_{\partial D^k} (f_j^k(x,t) - f_j^{k,*}(x,t))\hat{n}(\boldsymbol{M}^k)^{-1}\underline{l}(x) = \left(\frac{\lambda_j - |\lambda_j|}{2}\right)\hat{n}_r^- [[u]]_r - \left(\frac{\lambda_j + |\lambda_j|}{2}\right)\hat{n}_l^- [[u]]_l.$$

Here $\hat{n}_{r/l}^-$ denotes a vector pointing to the corresponding direction on the element's edge and $[[u]]$ indicates an upwind flux. λ_j corresponds to the diagonal elements of the matrix $\boldsymbol{\Lambda}$, as already introduced.

III. Validation and Computation Efficiency

To validate the approach, a RTD with a valence band energy, according to the schematic overview in Fig. 1, with a double barrier potential is chosen.

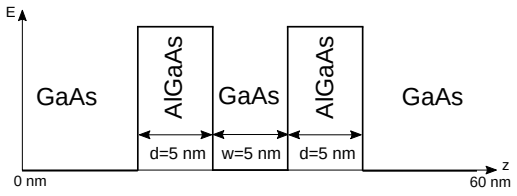


Fig. 1. Structure of a RTD with a valence band energy profile [4].

The density matrix is calculated for the thermal equilibrium. Additionally, a flatband case is assumed. The result for the absolute value of the statistical density matrix is shown in Fig. 2.

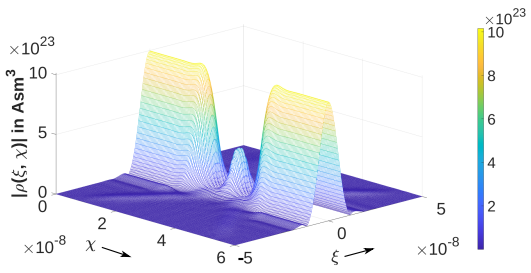


Fig. 2. Absolute value of the statistical density matrix $\boldsymbol{\rho}$.

The result is in good agreement with the result shown in [4]. However, when upwind fluxes are applied, an error occurs typically, as explained in [4], which leads to minor deviations compared to [4]. With a finer discretization in χ-direction the error converges to zero. Finally, the DG scheme is conceptually suited for solving the QLTE.

To demonstrate the computation efficiency, computation times of the DG-approach are compared to those of an FV-approach. As discussed, particularly the transient solution is of interest. To allow for an objective comparison, the average computation time for each time step for the transient solution will be calculated. The number of discretization elements has a major impact on the computation time. Accordingly, the number of N_ξ-elements will be varied for both algorithms between 50 and 350, in 50 block increments. All other parameters remain untouched.

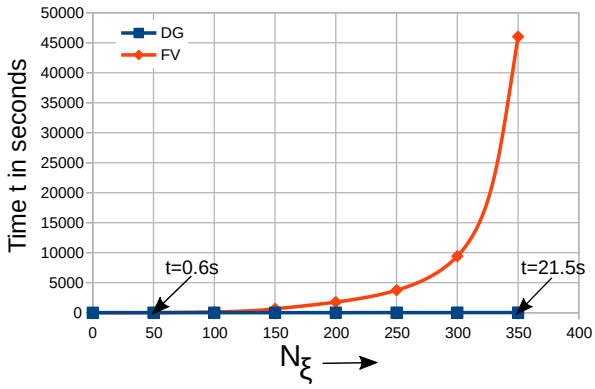


Fig. 3. Time comparison DG and FV for different ξ-elements

As can be concluded from Fig. 3, the computation time of the Finite Volume scheme shows an exponential-like increase depending on N_ξ, whereas the computation time of the DG scheme increases proportionally to the increase of N_ξ-elements. Finally, the DG scheme is shown to be more efficient than conventional methods enabling the analysis of larger scale problems.

References

[1] L. Schulz and D. Schulz, "Numerical Analysis of the Transient Behavior of the Non-Equilibrium Quantum Liouville Equation", in *IEEE Trans. on Nanotechnol.*, vol. 17, no. 6, pp. 1197–1205, Nov. 2018, 10.1109/TNANO.2018.2868972.

[2] K.-Y. Kim and B. Lee, "On the high order numerical calculation schemes for the wigner transport equation," *Solid-State Electronics*, vol. 43, no. 12, pp. 2243–2245, 1999. doi: 10.1016/S0038-1101(99)00168-9

[3] W. Frensley, "Wigner-function model of a resonant-tunneling semiconductor device," in *Phys. Rev. B*, vol. 36, no. 3, pp. 1570–1580, 1987, 10.1103/PhysRevB.36.1570.

[4] K. S. Khalid, L. Schulz, and D. Schulz, "Self-energy concept for the numerical solution of the liouville-von neumann equation," *IEEE Transactions on Nanotechnology*, vol. 16, no. 6, pp. 1053–1061, 2017. doi: 10.1109/TNANO.2017.2747622

Hybrid 2D/3D Mesh for Efficient Device Simulation of a Locally Tilted Vertical NAND String

Geon-Tae Jang and Sung-Min Hong
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology,
123 Cheomdan-gwagiro (Oryong-dong), Buk-gu, Gwangju, 61005, Republic of Korea
E-mail: smhong@gist.ac.kr

Abstract

In this work, we propose an efficient method to simulate a partially tilted vertical NAND string. The entire device is divided into several sub-devices. Each sub-device can have either a 2D mesh or a 3D one. The proposed method is applied to cases with and without structural deformation. It has been numerically demonstrated that the simulation becomes much more efficient with the hybrid 2D/3D mesh, while there is no loss of accuracy.

Introduction

Recently, as the demand for high storage NAND flash memory in various fields increases, major manufacturers are making efforts to increase the number of stacked layers in the vertical direction [1]. Channel holes are formed through a high aspect ratio etching process, and undesirable geometrical deformation occurs in this process [2]. In order to minimize the negative impact of structural deformation, most manufacturers have adopted the multi-stack process [3,4]. However, in the multi-stack process, a misalignment between two stacks may occur as shown in Fig. 1. Due to such a misalignment, the 3D NAND string is locally tilted. Performance of a 3D NAND Flash memory string can be predicted by the quasi-2D simulation using the cylindrical coordinate system. However, if the rotation axis offset occurs due to the misalignment mentioned above, the quasi-2D simulation is no longer available. In this case, it is appropriate to perform the full 3D simulation, but the full 3D simulation suffers heavily from a high computation burden.

In this work, we propose an efficient method to simulate partially deformed structures. A 3D mesh is used for a part with structural deformation from the rotationally symmetric structure. For all other parts away from the structural deformation, 2D meshes are adopted for efficiency. As an example, the proposed method is applied to a locally tilted vertical NAND string.

Device Structure and Simulation Methodology

An in-house device simulator has been newly developed from scratch to support a hybrid 2D/3D mesh. In order to show the feasibility of using a hybrid 2D/3D mesh, a cylindrical 3D device with two control gates and source/drain contacts is tested. Figure 2 shows the device structure used in the simulation. The silicon channel radius and insulator thickness are assumed to be 10 nm and 2 nm, respectively. Source/Drain doping is about 10^{20} cm^{-3} and an intrinsic channel is assumed. The entire device is divided into three sub-devices. Each sub-device can have a 2D mesh or a 3D one. The Poisson equation and two continuity equations are solved for these sub-devices, and as a result, the solution variables and the I-V characteristics can be obtained.

One of the most important things in this work is to connect adjacent sub-devices, and the method is as follows. First of all, we need to compute the interface points between adjacent sub-devices. In the case of a 2D/2D interface or a 3D/3D one, interface points can be found in a relatively simple way using vertex positions. In the case of a 2D/3D interface, a point in the 3D mesh is connected to a point in the 2D mesh by finding the radius from the position information. After that, for each sub-device, the Poisson and two continuity equations are implemented using its own mesh. Finally, the equations implemented in adjacent sub-devices are merged using the interface information calculated in advance.

Result and Discussion

When the entire device in Fig. 2 is divided into three sub-devices, 8 different cases occur in total. Among them, the electrostatic potential profile for three representative cases is shown in Fig. 3. Figure 4 shows the I-V characteristics calculated for 8 cases made with 3 sub-devices. The drain voltage is 0.1V and the lower control gate voltage (V_{CG_L}) is fixed as 1.0 V. Then, the drain current is calculated as a function of the upper control gate voltage (V_{CG_U}). For this rotationally symmetric structure, the quasi-2D/quasi-2D/quasi-2D case, which is calculated in the cylindrical coordinate system, yields the most accurate result. The 3D/3D/3D result is the most inaccurate result among all 8 cases, and the error is originated from a finite number of points along the angular direction. Figure 5 shows the elapsed time normalized to that of the quasi-2D/quasi-2D/quasi-2D case, when 1.0 V is applied to the upper control gate. It is confirmed that the elapsed time required to solve the coupled set of equations increases rapidly as more 3D meshes are introduced. Figure 6 shows a part of a vertical NAND string with a local tilt, originated from the misalignment between stacks. Our method can be applied even when an offset of the rotation axis occurs. Figure 7 shows the entire device used in the simulation. Although the 3D simulation should be performed for the tilted part, the quasi-2D simulation is possible for the parts away from the structural deformation. In the case of a misalignment between the upper and lower stacks, the electrostatic potential and the carrier densities may be rotationally asymmetric in the space near the offset. To alleviate such an asymmetry effect, the central sub-device includes also some buffers. Figure 8 shows the electrostatic potential calculated using the quasi-2D/3D/quasi-2D sub-devices when the offset is 0, 5, 10, or 15 nm. Figure 9 shows the drain current as a function of the rotation axis offset. The drain current decreases as the offset increases.

Conclusions

In conclusion, we have proposed and verified a method that enables an efficient simulation of a hybrid 2D/3D mesh without loss of accuracy.

References

[1] J. Choe, SISPAD, 2021.
[2] S.Park *et al*., VLSI, 2021.
[3] J. H. Kim *et al*., VLSI, 2021.
[4] H. Huh *et al*., ISSCC, 2020.
[5] J. Choe, FMS, 2019.

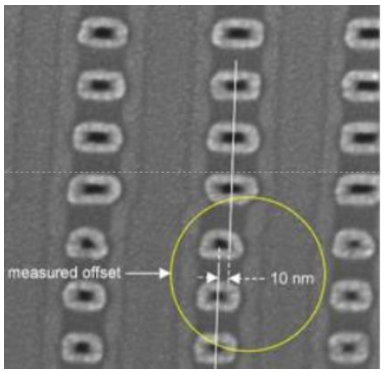


Figure 1. Misalignment between upper and lower stack [5]. A locally tilted 3D NAND string is observed.

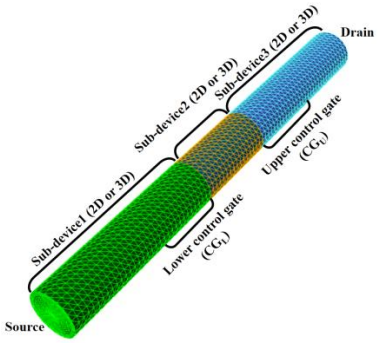


Figure 2. Simulated device structure. It has two control gates and source/drain contacts. The entire structure is divided into three sub-devices.

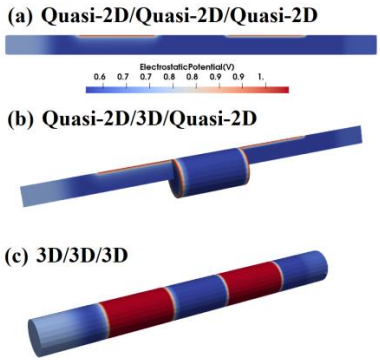


Figure 3. Electrostatic potential of three representative cases. (a) Quasi-2D/quasi-2D/quasi-2D, (b) quasi-2D/3D/quasi-2D, and (c) 3D/3D/3D.

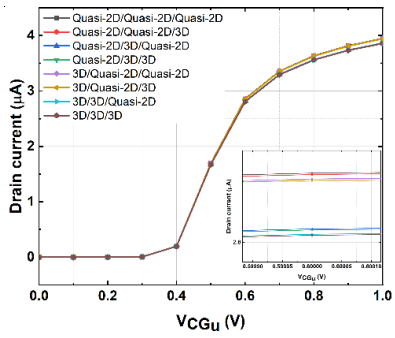


Figure 4. Calculated I-V characteristics of the eight cases. The applied drain voltage and the lower control gate voltage are 0.1 V and 1.0 V, respectively.

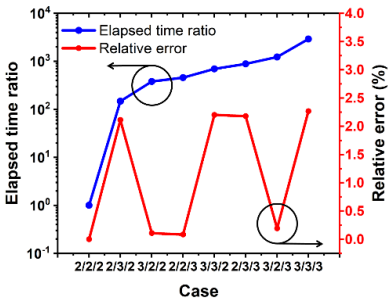


Figure 5. Elapsed time normalized to the quasi-2D/quasi-2D/quasi-2D case. 2 and 3 in x-axis represent the quasi-2D and the 3D, respectively. Relative error of the drain current is also shown.

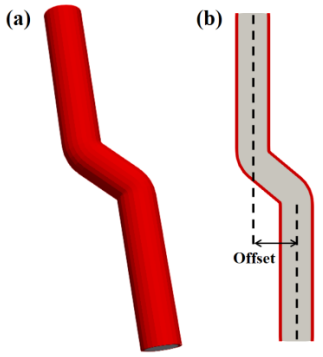


Figure 6. (a) 3D structure and its (b) cut plane of a device which has vertically stacked control gates with rotation axis misalignment.

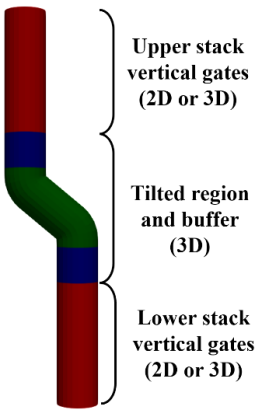


Figure 7. Simulated tilted device structure. In the case of the tilted device, some buffers are added for the central 3D region.

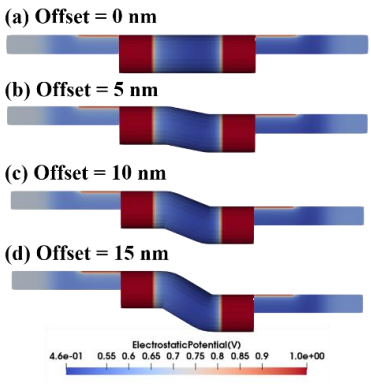


Figure 8. Electrostatic potential obtained by the nonlinear Poisson equation. The quasi-2D/3D/quasi-2D sub-devices are sued. The offset varies from 0 nm to 15 nm.

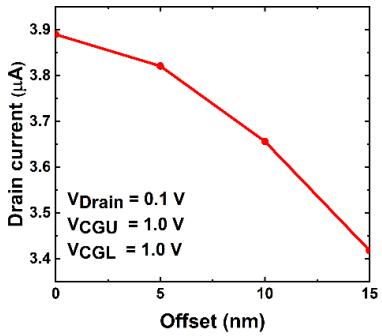


Figure 9. Drain current as a function of the rotation axis offset. The applied voltage of V_{Drain}, V_{CG_L}, and V_{CG_U} is 0.1 V, 1.0 V, and 1.0 V, respectively.

Stacking devices in a vertical nanowire, a feasible option to implement smaller ICs

E. Amat, A. del Moral, J. Bausells, *Senior Member, IEEE* and F. Perez-Murano
Institute of Microelectronics of Barcelona (IMB-CNM, CSIC), 08193 Bellaterra, Spain.
esteve.amat@imb-cnm.csic.es

Abstract — **Devices based on 3D configuration by using a vertical topology are being considered as the next step to improve electronic devices and circuits performance. A new IC distribution has been proposed by stacking devices to implement a CMOS inverter using Sentaurus 3D TCAD. We have also explored their feasibility to implement a 5-stages ring oscillator circuit.**

I. Introduction

The continuous MOSFET's scale down is looking for the improvement of the throughput of integrated circuits (IC). One of the most relevant breakthroughs for next generation of electronic devices is its implementation in vertical topology. Beyond 10 nm nodes, a fully gate-all-around (GAA) configuration is necessary to improve the device throughput to obtain a better control over channel conduction; in this case nanowire-based FET (NW) is a promising device candidate. Its fully 3D configuration implies a relevant mitigation of the short channel effects (SCE) [1], which in turn implies a significant reduction in device variability. Additionally, vertical topologies (vNW) have demonstrated larger scalability than lateral ones [2], outperforming FinFET architectures in terms of speed and power consumption [3]. These devices are less restricted on gate length and spacer thickness, as their vertical orientation allows certain design relaxation [4]. Vertical structures are gaining relevance because they are very interesting to implement low-power circuits and bio-sensors [5]. In this context, IRDS 2021 forecast [6] states that stacking p- and n-devices is a valid proposal for the future of electronic circuits. Monolithic 3D approximation explores by stacking planar MOSFETs in vertical to reduce both area foot-print and power consumption [7]. In this sense, we explore an original proposal to implement an IC by stacking both devices (p and n) in the same vertical nanowire structure, reducing the area footprint and enhancing circuit throughput.

II. Simulation Framework

The whole circuit structure and its behavior have been simulated by using Sentaurus 3D TCAD software [8], which allows us to analyze different relevant factors; e.g. structure, physical aspects (electric field, current flow) and electrical behavior. As prove of concept of our circuit implementation proposal, we have simulated a CMOS inverter, due to its simplicity. Fig. 1.a presents a sketch of our circuit proposal topology, by stacking the pMOS device over the nMOS one with the output terminal between both devices. The source terminal is connected to the bulk and at ground level. The supply voltage (V_{DD}) of the inverter is connected at the top of the silicon pillar. For simulation simplicity, all design parameters of the circuit have the same starting value stated at 30 nm. The core IC structure consists of a vertical silicon nano-pillar of 30 nm of diameter. A gate dielectric layer with an equivalent oxide thickness (EOT) of 2 nm has been selected to cover the region dedicated to the device channel. For the ground contact of the CMOS inverter (GND) we deposit a 10 nm Nikel layer. For the metal layers surrounding the pillar as a GAA device, i.e. output and inputs, are based on Tungsten (W). All the interconnections and contacts are implemented by using Aluminum layers. Note that taking into account that the dimensions of the structure are always at the range of tens of nanometers, the Hydrodynamic and Quantum Potential models have been used to achieve a more realistic behavioral study.

This CMOS inverter proof-of-concept study is based on the simulation of conventional MOSFET device, based on p and n junction regions. The pillar is patterned from a bulk p-type silicon substrate, boron-doped with a doping level of $1.0x10^{15}$ at cm^{-3}. Fig. 1.b presents a cross-section with the different regions of the stacked devices. In between the pMOS and nMOS devices, in the middle there is the output terminal. It is worth to mention that to isolate both devices there is a low doped ($1x10^{15}$ at cm^{-3}) middle region (LDMR) of 10 nm to ensure a negligible leakage through both devices. For simplicity, both electrodes are equally phosphorous-doped with $S/D_{implant}$ = $1x10^{20}$ at cm^{-3} with a constant doping along the source and drain regions. The p- and nMOS devices should be

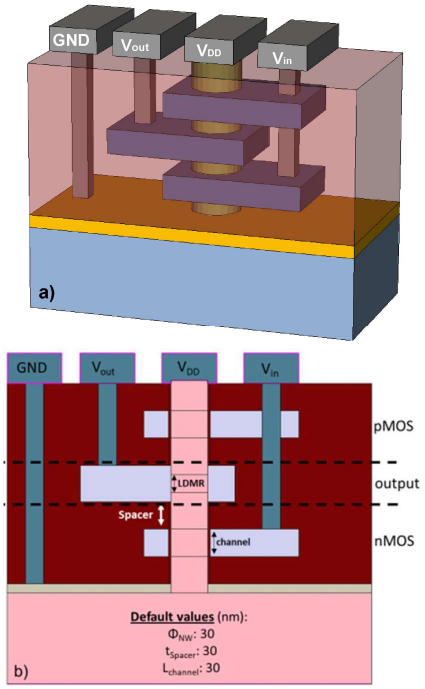


Figure 1. a) 3D TCAD structure of the simulated CMOS inverter schematic. b) Circuit cross-section of the stacked CMOS inverter proposal.

well balanced to allow both devices operate in all regimes; for this reason, the nMOS channel doping is $1x10^{18}$ at cm^{-3}. To extend the application of this device fabrication strategy, a 5-stage ring oscillator (ROSC) will be simulated as well. Their structure is based on the reference inverter reproduced 5 times to achieve the final 5-stage ROSC. Its feasibility will be analysed by considering the output frequency as a main parameter.

III. Feasability of the IC Stacked Proposal

First, seeking to verify the functionality of our stacked inverter proposal, we have analyzed its suitability in transient regime. So, a pulsed signal is applied to the input terminal (V_{in}), and we expect the opposite signal at the output terminal (V_{out}). Thus, Fig. 2.a shows that the behavior of our IC proposal is correct and the expected inversion is obtained. Afterwards, the stacked IC proposal has been evaluated in DC regime to obtain the voltage characteristic curve (VTC). Fig. 2.b shows the inverter performance in DC regime, by proving the well design of both devices and the good behavior as digital gate. Moreover, we have analyzed its suitability at a wide V_{DD} range, from 0.2 to 1.2V; and our proposal presents a promising behavior even at ultra-low V_{DD} in corroboration with [9], by proving that vNW presents a relevant applicability at this supply levels. This performance can highlight the usefulness of this proposal into low power consumption environments, interesting for the Internet of Things (IoT) and bio-sensors.

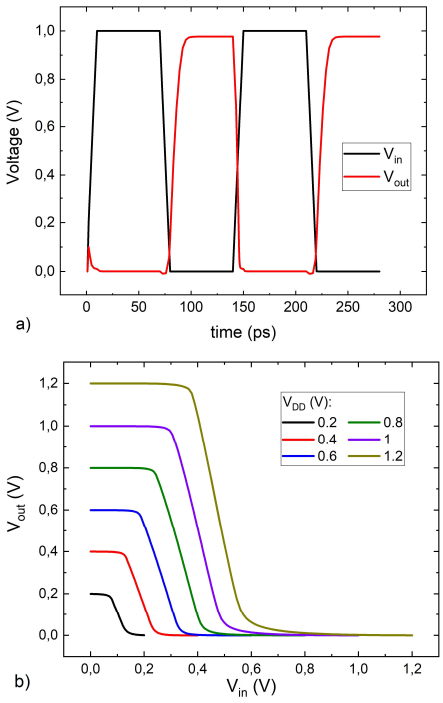


Figure 2. Study of the feasability of the CMOS inverter based on the stacked IC implementation proposal, by studing it in transient (a) and DC (b) regimes.

IV. Extension of the Usefulness of Stacked Approach

A clear extension of our proposal implementation is its use as a ring oscillator circuit (ROSC), where a chain of CMOS inverters are connected in a row. In terms of manufacturing process, it should not involve a relevant difficulty for our stacked proposal implementation, because this chain of inverters could be manufactured in an array configuration, what involves an easy fabrication process. Fig. 3.a presents a cross-section of the proposed 5-stages ROSC structure based on stacked configuration simulated by 3D TCAD. To implement this 5-stages ROSC we have used 5 nanopillars for each stacked CMOS inverter, instead of a single vNW in order to simplify the IC structure. Note that stacking ten devices in a single nanopillar could involve a relevant threat in terms of structure stability, as it would need to make all the process with an almost 1 µm pillar height. However, although this proposal is not so aggressive in terms of area foot-print, it would entail a relevant area reduction in front the classic approach. Fig. 3.b presents the output signal of the reference ROSC proposal, and a proper oscillating output signal is obtained with a frequency about 20GHz. It is worth to mention that by modifying the device dimensions, we can tune the output frequency in our convenience.

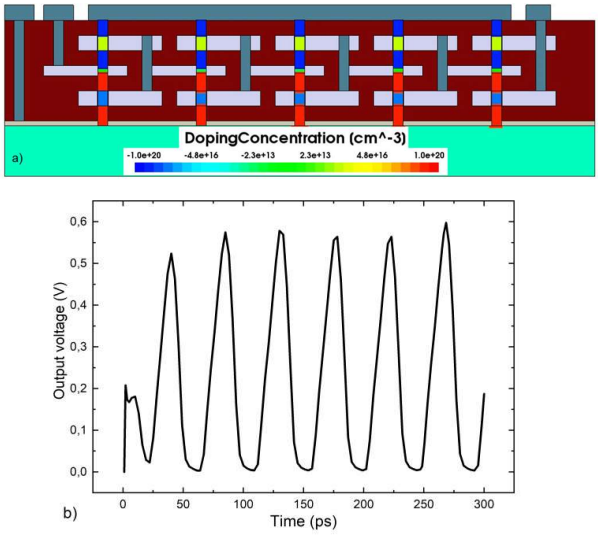


Figure 3. Extension of the stacked IC implementation proposal by simulating a 5-stages ROSC cross-section (a). Simulation of the impact of the b) pillar diameter.

V. Conclusions

We present a suitable novel approach to implement CMOS integrated circuits. The basis of our proposal is to stack two devices in a vertical nanopillar. To prove their feasibility, we have simulated a CMOS inverter as proof-of-concept. This proposal entails a relevant reduction in area foot-print by exploring the use of extreme 3D topology, in front of using the conventional proposal of two devices in parallel. We have analyzed the structure behavior by using 3D TCAD Sentaurus numerical simulations. To extend the usefulness of this new manufacturing strategy we have also explored the implementation of a more complex circuit, a 5-stages ROSC. From the simulations we have observed a proper behavior of this circuit. This new manufacturing proposal exposed in this contribution can entail a relevant benefit for the semiconductor industry in terms of area foot-print reduction and throughput.

References

[1] X. Zhao *et al.*, *IEDM*, pp. 28.4.1-28.4.4, 2013.
[2] D. Yakimets *et al.*, *IEEE TED*, vol. 62 **(5)**, pp. 1433–1439, 2015.
[3] T. Huynh-Bao *et al.*, *Proc. SPIE*, vol. 9781, pp. 978102–978112, 2016.
[4] D. Yakimets *et al.*, *Device Res. Conf.*, pp. 133–134, 2014.
[5] P. Saha *et al.*, *Solid. State. Electron.*, vol. 161, p. 107637, 2019.
[6] IEEE, "International Roadmap for Devices and Systems (IRDS)," 2021.
[7] A. B. Sachid et al., *Appl. Phys. Lett.*, vol. 111 **(22)**, p. 222101, 2017.
[8] S.D.U. Guide, "Sentaurus workbench user guide version K-2018.06."2018.
[9] C. Pan *et al. IEEE TED*, vol. 62 (10), pp. 3125–3132, 2015.

A comprehensive on-current variability Pelgrom-based model for FinFET, NWFET and NSFET transistors

Julian G. Fernandez*, Natalia Seoane, Enrique Comesaña and Antonio García-Loureiro
CiTIUS, University of Santiago de Compostela, Spain (*e-mail: julian.garcia.fernandez2@usc.es)

THE scaling of semiconductor devices is essential for the progress of the electronic industry, but their small size requirements make them more vulnerable to different sources of variability. The metal grain granularity (MGG) and the line edge roughness (LER) are two of the variability sources that have the greatest impact on state-of-the-art transistor architectures [1]: nanowire (NW) FETs, FinFETs, and nanosheet (NS) FETs. TCAD is commonly used in variability studies since, to obtain statistical significance, we require the analysis of a large number of samples. The computational cost of on-region variability studies can be prohibitive because either quantum-corrected Monte Carlo, or full quantum simulations are necessary to properly capture non-equilibrium transport effects. For this reason, new strategies are needed to reduce computational times of statistical studies [2], [3].

In this work, we propose a novel Pelgrom-based predictive (PBP) model to estimate the impact of variability on the on-current standard deviation (σI_{on}) due to MGG and LER variabilities. The Pelgrom's Law [4] states that the standard deviation of a figure of merit is proportional to the inverse square root of the effective gate area, defined as the product between the gate length (L) and the effective gate perimeter (W) [5], [6]: $\sigma I_{on} = \frac{A_i}{\sqrt{LW}}$. The A_i is the on-current matching factor, which is determined by the contributions of all possible sources of transistor variations [7]. We have developed two models based on Pelgrom's Law, one for MGG and other for LER variability, applying them to the three architectures shown in Fig. 1. To validate the PBP model we have used data published by different authors [8], [9], [10], [11] combined with simulations that were done for this study using quantum corrected Monte Carlo methodology in VENDES [12] tool.

MGG depends on the average metal grain size (GS), which is determined by the annealing temperature used in the gate deposition process [13] (see Fig. 1(c)). To develop the PBP model for MGG variability, we assume that MGG is the sole contributor to transistor variations, and therefore $A_i = A_{mgg}$. Also, we assume that σI_{on} must depend linearly on the GS [8], [10], leading to the PBP model for MGG:

$$\sigma I_{on} = A_{mgg}\frac{1}{\sqrt{LW}} = \theta_{mgg}\frac{GS}{\sqrt{LW}} \tag{1}$$

We define the θ_{mgg} as the on-current mismatch for MGG, which is a technological parameter that states how the variability impacts on a certain architecture, and once is determined, the σI_{on} could be predicted for any device dimensions at any GS. The PBP model plots for MGG are shown for the three architectures in Fig. 2(a)-(c). The minimum sample size used for each set of simulations is 300 in order to obtain statistical significance. The black line is the reference line that denotes the fitting of the model, where the error bars correspond to 10% of the relative error with respect to the simulated data, a tolerance that we consider acceptable. As can be seen, the majority of the predictions are inside the 10% margin. Table I presents the predicted (σI^P_{on}) and simulated (σI^S_{on}) on-current standard deviations together with the relative error (σ_r) between them, the GS, the θ_{mgg}, and dimensional information of the devices (LW). The estimations outside the 10% margin of error are highlighted in bold in the table, and are due to the saturation of the standard deviation in devices with small LW and large GS (i.e. NWFET L=10nm with GS=7 nm and 10 nm), this phenomenon is described in [14].

LER variability is reproduced in TCAD studies using the Fourier transform of the Gaussian spectra [15] (see Fig. 1(b)), depending on two parameters: the root mean square (Δ, depth of the roughness) and the correlation length (Λ, propagation of the roughness). In this case, the LER is the sole contributor to variability $A_i = A_{ler}$. Also, we assume that σI_{on} increases linearly with Δ, depending on the square root of the product between the device width (w) and a function of Λ ($f(\Lambda) = \Lambda \cdot (1 - e^{-\sqrt{\frac{L}{\Lambda}}})$). Thus, the PBP model for LER will be as follows:

$$\sigma I_{on} = A_{ler}\frac{1}{\sqrt{LW}} = \theta_{ler} \cdot \Delta\sqrt{\frac{f(\Lambda)\cdot w}{LW}} \tag{2}$$

The θ_{ler} will be the on-current mismatch for LER, is a parameter that states the dependence of variability due to an architecture, and once is fixed, the estimation of σI_{on} could be done at any dimensions (L, W, w), Δ, or Λ. The PBP model plots are shown in Fig. 2(d)-(e) for NWFET and FinFET due to LER induced variability, respectively. Also, in Fig. 2(f) we show a comparison between the simulated and predicted σI_{on} due to LER variability, together with the shaded region where the deviations are lower than 10% from the expected value for the 22 devices studied. We can see a good match between the predicted and the simulated data for different values of Δ, Λ, different dimensions, and architectures.

In conclusion, we have presented an on-current variability prediction model for MGG and LER based on Pelgrom's Law. It has been tested for three state-of-the-art architectures, with prediction errors lower than 10% in the 96% of the cases. Therefore, the PBP model is a simple, fast, and reliable strategy to estimate the impact on the transistor performance of a certain source of variability. This model could be useful to predict the impact of variability on future technology nodes.

REFERENCES

[1] N. Seoane *et al.* doi: 10.1109/LED.2021.3109586
[2] H. Carrillo-Nuñez *et al.* doi: 10.1109/LED.2019.2931839
[3] G. Indalecio *et al.* doi: 10.1109/TED.2017.2670060
[4] M. J. M. Pelgrom *et al.* doi: 10.1109/JSSC.1989.572629
[5] A. Sheikholeslami. doi: 10.1109/MSSC.2014.2369331
[6] N. Lu *et al.* doi: 10.1109/IEDM.2014.7047174
[7] P. Stolk *et al.* doi: 10.1109/16.711362
[8] N. Seoane *et al.* doi: 10.1109/TED.2016.2516921
[9] W. Sung *et al.* doi: 10.1109/JEDS.2020.3046608
[10] D. Nagy *et al.* doi: 10.1109/jeds.2018.2804383
[11] D. Nagy *et al.* doi: 10.1109/ACCESS.2020.2980925
[12] N. Seoane *et al.* doi: 10.3390/ma12152391
[13] H. Dadgour *et al.* doi: 10.1109/IEDM.2008.4796792
[14] J. G. Fernandez *et al.* doi: 10.1109/SISPAD54002.2021.9592600
[15] A. Asenov *et al.* doi: 10.1109/TED.2003.813457

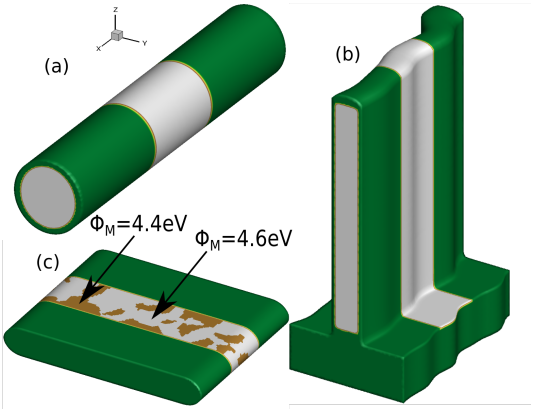


Figure 1: Scheme of a (a) NWFET not affected by any variability source, (b) FinFET affected by LER variability, and (c) NSFET affected by TiN MGG with two gate work-functions (ϕ_M).

Table I: Predicted (σI^P_{on}) vs. simulated (σI^S_{on}) on-current standard deviations, at different GS for each architecture. In the table we also list the on-current mismatch θ_{mgg}, the gate effective gate area (LW), and the relative error σ_r. The simulations done for this work are referenced with *.

	GS [nm]	σI^P_{on} [A/m]	σI^S_{on} [A/m]	$\theta_{mgg} = 120\ nA/nm$ σ_r	L×W [nm^2]	Ref.
NWFET	3	23.8	25.0	-4.8%	10.0×22.8	[10]
		11.5	12.5	-8.0%	22.0×44.9	[10]
	5	39.7	42.5	-6.6%	10.0×22.8	[10]
		19.1	19.0	+0.5%	10.0×35.2	[9]
		32.0	33.0	-5.3%	22.0×44.9	[10]
	7	55.6	47.7	**+16.6%**	10.0×22.8	[10]
		26.7	26.3	+1.5%	22.0×44.9	[10]
	10	79.5	67.7	**+17.4%**	10.0×22.8	[10]
		38.2	42.4	-9.9%	22.0×44.9	[10]
	GS [nm]	σI^P_{on} [A/m]	σI^S_{on} [A/m]	$\theta_{mgg} = 192\ nA/nm$ σ_r	L×W [nm^2]	Ref.
FinFET	3	14.9	15.0	-0.7%	25.0×60.0	*
		16.3	16.3	+0.0%	12.0×105.0	*
	5	52.7	58.6	-8.4%	10.7×30.0	[8]
		24.9	24.9	+0.0%	25.0×60.0	*
		27.1	27.1	+0.0%	12.0×105.0	*
	7	75.2	75.9	-0.9%	10.7×30.0	[8]
		34.8	36.4	-4.5%	25.0×60.0	*
		38.0	35.3	+7.6%	12.0×105.0	*
	10	107.4	105.9	+1.4%	10.7×30.0	[8]
		49.7	47.1	+5.5%	25.0×60.0	*
		54.2	53.3	+1.7%	12.0×105.0	*
	GS [nm]	σI^P_{on} [A/m]	σI^S_{on} [A/m]	$\theta_{mgg} = 191\ nA/nm$ σ_r	L×W [nm^2]	Ref.
NSFET	3	15.8	15.2	+3.9%	12.0×110.0	[11]
		10.5	11.4	-7.9%	18.0×165.3	*
	5	26.3	25.0	+5.2%	12.0×110.0	[11]
		17.5	16.5	+6.1%	18.0×165.3	*
	7	36.8	36.5	+1.0%	12.0×110.0	[11]
		24.5	24.8	-1.2%	18.0×165.3	*
	10	52.6	52.3	+0.7%	12.0×110.0	[11]
		35.0	32.8	+6.7%	18.0×165.3	*

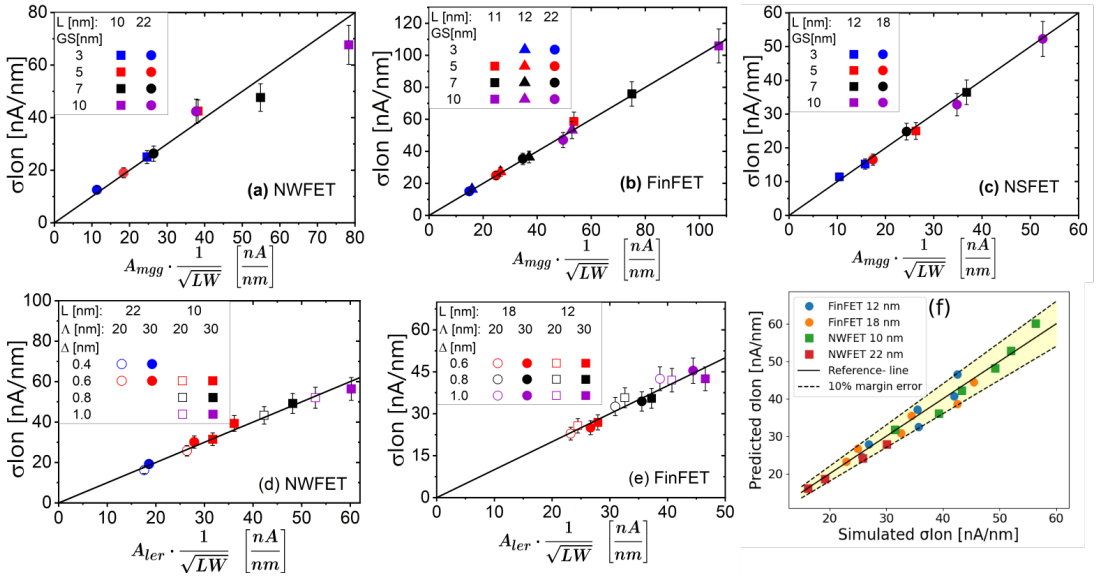


Figure 2: Pelgrom-based predictive (PBP) plots for MGG variability for three state-of-the-art different arquitectures: (a) NWFET, (b) FinFET, and (c) NSFET. PBP plots for LER variability for (d) NWFET, (e) FinFET. Also, in (f) is shown a comparison between data predicted with the PBP model and the simulated data due to LER. The gate length L of the devices, together with the variability parameters (GS for MGG, Δ and Λ for LER), are also shown.

Forked Contact and Dynamically-Doped Nanosheets to Enhance Si and 2D Materials Device at the limit of Scaling

Aryan Afzalian, Zubair Ahmed and Julien Ryckaert
imec, Leuven, Belgium, aryan.afzalian@imec.be

***Abstract*—We propose a novel Forked-Contacts, Dynamically-Doped Multigate transistor as ultimate scaling booster for both Si and 2D materials in aggressively-scaled nanosheet devices. Using accurate dissipative DFT-NEGF atomistic-simulation fundamentals and cell layout extrinsics, we demonstrate superior and optimal device characteristics and invertor energy - delays down to sub-30-nm pitches, i.e., a 10 nm scaling boost compared to the nanosheet MOSFET references, regardless of the material system used. This gain is linked to a more compact architecture but does not change the material-specific fundamental gate-length limit that we also assess here. By switching from Si to 2D materials, however, an additional 5 nm reduction in gate length scaling could be enabled. Keywords: CMOS, CGP scaling, Si, 2D materials, DFT-NEGF, ab-initio, nanosheet, dynamic-doping**

Introduction

The Dynamically-Doped (D2) Field-Effect Transistor is a novel device architecture that scales better than its MOSFET nanosheet (*NS*) counterpart [1], owing to the suppression of ungated extensions (spacers) from the device Contacted Gate-pitch (CGP) equation [1,2] and Fig. 1a. What used to be the *NS* chemically doped extensions are now electrically and dynamically-doped by the gate, i.e., a part of the channel. Hence, for a given CGP, the channel length L in the D2FET is twice the spacer length (L_{SPACER}) longer than L of a standard MOSFET, as it benefits from the full distance between the source (S) and drain (D) contact pads. The gate length L_G value could even be larger than L, if the gate would overlap over the contact region of length L_C (Fig. 1). For a single-gate (SG) single-sheet device, this can simply be enabled by having the gate contact on the side opposite to the contacts, i.e., using for instance a top-contact and an individually back-gated transistor [1]. To enable a D2 tri-gate with stacked sheets, however, we propose here a doubled forked structure (E2), where the sheets are connected to a forked gate on one side and to forked S & D contacts on the other side (Fig. 1). Our simulation results show, as expected, that such a multigate E2D2 architecture enables a better electrostatic control and improved drive current at scaled CGP, especially for Si where the film thickness can be relaxed, compared to the SG-D2 transistor. We report here on the impact of the multigate E2D2 architecture innovation on intrinsic-device and loaded-invertor performance, when pitch is scaled well below 30 nm using accurate dissipative DFT-NEGF atomistic-simulation fundamentals and cell-layout extrinsics. The E2D2 architecture is benchmarked to *NS* MOSFETs using both Si and 2 emerging 2D transition metal dichalcogenide (TMD) monolayer (1ML) materials – one, WS_2, with predicted fundamental drive similar to that of Si, the other, HfS_2, featuring an enhanced fundamental drive current [1] – as test vehicles.

Methods

Current – Voltage ($I_D(V_G)$, Fig. 2d inset) and intrinsic device capacitances (C_{Gi}) (Fig. 2b) for Si and 2D TMD E2D2 and standard *NS* references are simulated using our first-principle atomistic NEGF solver ATOMOS, including electron-phonon scattering [1,2]. From these simulations the intrinsic single-sheet device fundamental performance vs. CGP can be assessed (Fig. 2). For each CGP, a full device optimization is made including film thickness (t_S) scaling for Si and extension doping for the *NS*. The detrimental impact of quantum confinement, including a ballistic ratio decrease due to an increase of the electron-phonon wave-function overlap [3,4], dark space [5], and source-to-drain direct tunneling are included in our quantum transport solver. For computing stacked-invertor energy-delay products (Fig. 3), the extracted extrinsic capacitance of the cell layout C_{cell} and the backend-of-line load, C_{BK} are used (Fig. 1). C_{cell} values are reported in Fig. 4a. The number of stacked sheets (n_S) used is computed to allow a total stack height of 60 nm for all devices. n_S is the same for E2D2 and *NS* of a same material (n_S = 4 for Si and 5 for the TMDs owing to their 1ML thickness of about 0.6 nm [1]). The available width for a single sheet, W, in our 5-track E2D2 layout cell is 12 nm. The standard *NS* layout is described in [6] and W is 12 nm as well.

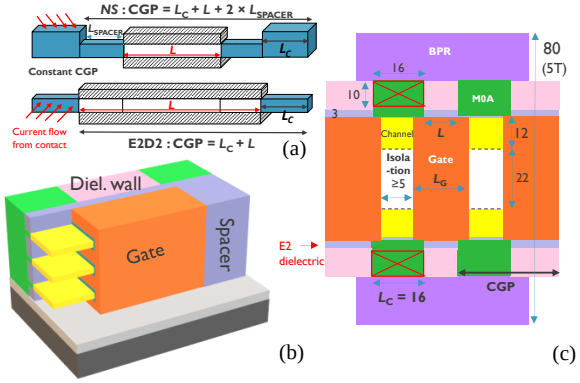


Fig. 1 E2D2 device structure. a) Side view schematic of a single-sheet multigate conventional *NS* (Top) and E2D2 transistor (Bottom) with same CGP. b) 3D view showing the doubled forked (E2) structure, c) cell layout of the 5-track (cell height = 80 nm) E2D2 invertor cell with buried power rail (BPR). The technological dimensions, we assumed for this study are indicated in the figure. We assumed $L_G = L$. L_C = 16 nm, L_{SPACER} = 5 nm. The width of an individual sheet is W = 12 nm. The P/N separation is 22nm. For the gate oxide, we assumed a 2 nm hafnium oxide with ε_R = 15.6. The gate stack metal thickness is 6 nm. The channel (white region in Fig. 1a) is intrinsic. The contact regions (in blue in Fig. 1a) are doped. n_S = 4 for Si and 5 for 2D.

Results

Owing to its 10 nm extended gate length at same CGP, the E2D2 SS and, hence, I_{ON} at fixed I_{OFF} are superior compared to that of the *NS* as CGP is scaled below 30 nm for all materials. The E2D2 C_{Gi} are however larger at fixed CGP, the net effect being that the E2D2 optimal intrinsic delay is comparable to that of its *NS* counterpart, but shifted towards smaller CGPs by about 10 nm, i.e., $2\times L_{SPACER}$ (Fig. 2). Hence the E2D2 architecture enables a significant scaling boost. For Si, the optimal *NS* and E2D2 delays are obtained at CGP = 36, and 26 nm, respectively, i.e., L_G = 10 nm and t_S = 3 nm in both cases. For the 2D materials, a further 5 nm scaling boost is observed, and optimal delays are achieved at CGP = 31 and 21 nm for the *NS* and E2D2 respectively, corresponding to L_G = 5 nm in both cases. For the TMDs the 21 nm E2D2 CGP corresponds to the case where CGP is only limited by the contacts (L_C and the minimum isolation spacing, IS, required to separate subsequent pads, assuming IS = L_{SPACER} (Fig. 1)), hence ultimate gate scaling has been achieved. Further CGP reduction may still be achieved by scaling the contacts.

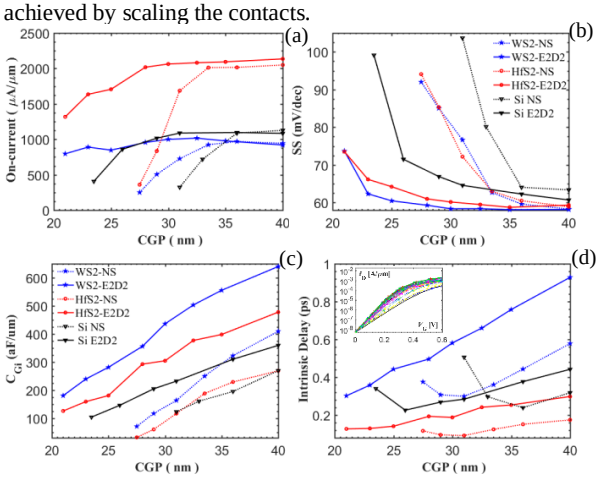


Fig. 2 a) On-current (I_{ON}) at fixed I_{OFF}, b) subthreshold slope (*SS*), c) intrinsic gate capacitance (C_{Gi}) and d) intrinsic delay - assuming an effective current i_{eff} = 0.45 × I_{ON} – vs. CGP for Si and for WS_2 and HfS_2 2D monolayers *NS* and E2D2 architectures from ab-initio - NEGF transport simulations (the trace of the $I_D(V_G)$ characteristics are in the inset) [1]. For the *NS*, L_G = CGP - L_C - $2\times L_{SPACER}$, while for E2D2 L_G = CGP - L_C. L_C = 16 nm, L_{SPACER} = 5 nm. I_{ON} is normalized by the gate perimeter. I_{OFF} = 5 nA/μm. V_{DD} = 0.6V.

Next, we investigate the switching energy *vs.* delay (EDP) of high-performance stacked E2D2 and *NS* loaded inverters for different CGPs at various V_{DD} (Fig. 3). For HfS_2 E2D2 and *NS* invertors, the optimal EDP is achieved at CGP = 21, L_G = 5 nm and CGP = 31, L_G = 5 nm respectively. We obtain a similar result for the WS_2 case (not shown here). For Si E2D2 and *NS* invertors, the optimal EDP is achieved at CGP = 26, L_G = 10 nm and CGP = 36, L_G = 10 nm respectively. Any further attempt to scale CGP by scaling L_G beyond this optimal value results in significant performance reduction (for the TMD E2D2 it is simply not possible to further scale CGP with L_G scaling). These results further confirm the 10 nm improved scalability, we obtained from the intrinsic device performance and delays (Fig. 2). The E2D2-invertor improved EDP performance compared to that of the *NS* is mostly linked to the reduced C_{BK} owing to CGP scaling.

This is confirmed in Fig. 4, where the loaded-invertor EDPs are shown for the 3 material cases at their optimal CGP and L_G values with and without C_{BK} included in the load. Regardless of the material system used, without C_{BK} (Fig. 4.a), the E2D2 performance are similar to that of their *NS* counterparts, while they are enhanced when C_{BK} is included (Fig. 4.b) (the E2D2 and *NS* devices also share the same optimal L_G of 10 nm for Si and 5 nm for the TMDs).

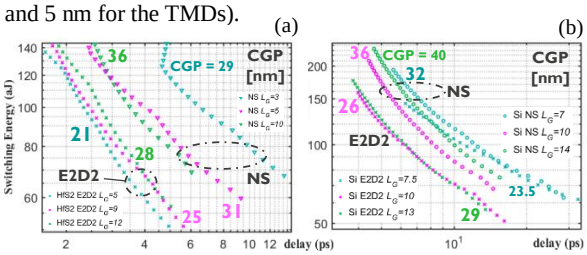


Fig. 3 Switching energy *vs.* delay (EDP) of high-performance stacked E2D2 and *NS* inverter cells for different CGP and L_G, as indicated in the figures, at various V_{DD} (0.4V to 0.7V). The devices are made of a) 1ML-HfS_2 with n_S = 5 sheets/device, b) Si with n_S = 4 sheets/device and optimized Si thickness t_S ranging from 3 to 5 nm. The inverters are loaded with the cell layout capacitances C_{Cell} and a 50 CGP-long metal line with capacitance C_{BK} = 198 aF/μm [7]. I_{OFF} = 5 nA/μm.

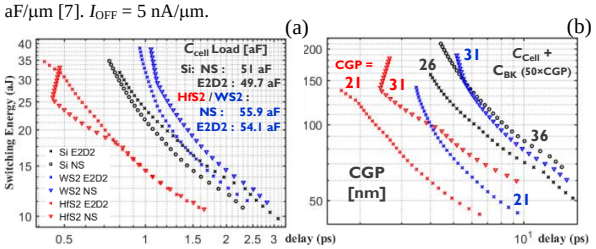


Fig. 4 EDP of the Si, WS_2 and HfS_2 stacked E2D2 and *NS* inverter cells at optimal CGP as indicated in Fig. 4b, at various V_{DD} (0.4V to 0.7V). The inverters are loaded with a) C_{cell} only, its value for each invertor case is indicated in the figure, b) C_{cell} and a 50 CGP-long metal line C_{BK}. I_{OFF} = 5 nA/μm.

Conclusions

We proposed a compact E2D2 multigate architecture that enables sub-30-nm CGP, i.e., an improved $2\times L_{SPACER}$ pitch scaling, compared to a *NS* reference, owing to the suppression of ungated extensions from the CGP equation. This E2D2 scaling benefits were measured in term of similar intrinsic performance and optimal delay but at a 10 nm reduced CGP. A similar conclusion was found comparing E2D2 and *NS* stacked-invertor cells. For backend-loaded invertors, the E2D2 EDP performance is further enhanced due to CGP and, hence, C_{BK} reduction. Similar relative benefits were observed regardless of the material system used. Compared to Si, a mature 2D material technology could potentially further enable an extra 5-nm CGP scaling boost, enabled by a smaller L_G both for the E2D2 and *NS* architectures, with same or improved performance, if respectively WS_2, a material with a fundamental drive similar to Si, or HfS_2, a higher mobility material, were used.

References

[1] A. Afzalian, “Ab initio perspective of ultra-scaled CMOS from 2d material fundamentals to dynamically doped transistors”, npj 2D Materials and Applications, vol. 5, no. 1, 2021.

[2] A. Afzalian, E. Akhoundi G. Gaddemane, R. Duflou and M. Houssa, “Advanced DFT–NEGF Transport Techniques for Novel 2-D Material and Device Exploration Including HfS2/WSe2 van der Waals Heterojunction TFET and WTe2/WS2 Metal/Semiconductor Contact”, in IEEE Transactions on Electron Devices, vol. 68, no. 11, pp. 5372-5379, Nov. 2021, doi: 10.1109/TED.2021.3078412.

[3] A. Afzalian, “Ultimate FDSOI multi-gate MOSFETs and multi-barrier boosted Gate Modulated Resonant tunneling-FETs for a new high-performance low-power paradigm”, Nano-Semiconductors: Devices and Technology, Kris Iniewski (Ed.) (CRC Press, Boca Raton 2012). https://doi.org/10.1201/9781315217468.

[4] A. Afzalian. "Computationally efficient self-consistent born approximation treatments of phonon scattering for coupled-mode space non-equilibrium Green’s function." *Journal of Applied Physics*, vol. 110, no. 9, p. 094517, 2011.

[5] Afzalian, A., Lee, C., Dehdashti Akhavan, N., Yan, R., Ferain, I. & Colinge, J. “Quantum Confinement Effects in Capacitance Behavior of Multigate Silicon Nanowire MOSFETs”, IEEE Trans. Nanotech. 10, 300-309 (2011).

[6] Z. Ahmed et al., Introducing 2D-FETs in Device Scaling Roadmap using DTCO, IEDM, 2020, pp. 22.5.1.

[7] https://irds.ieee.org/ editions/2018

On the Switching Limits of Top-Gated Carbon Nanotube Field-Effect Transistors

A. Sanchez-Soares*, C. Gilardi†, Q. Lin†, T. Kelly*, S.-K. Su‡, G. Fagas‖, J.C. Greer§, G. Pitner¶, E. Chen‡
*EOLAS Designs, Grenagh, Co. Cork, Republic of Ireland. E-mail: alfonso.sanchez@eolasdesigns.com
†Stanford University, Stanford, CA, USA. ‡Corporate Research, TSMC, Hsinchu, Taiwan.
§University of Nottingham Ningbo China, Ningbo, China. ¶Corporate Research, TSMC, San Jose, CA, USA.
‖ Tyndall National Institute, University College Cork, Cork, Ireland.

***Abstract*—The performance limits of carbon nanotube field-effect transistors (CNFETs) based on a recently reported process are studied by computational techniques at temperatures between 300K and down to 4K. The impact of band-to-band tunneling (BTBT) and source-to-drain tunneling (SDT) is examined for devices with varying CNT dimensions through the use of simulations based on the non-equilibrium Green's function (NEGF) formalism. Additionally, the case of chemical doping is analysed in contrast to electrostatic doping recently reported for test structures.**

I. Introduction

Logic transistors based on carbon nanotubes are promising candidates for increasing densities and power efficiency over Si-based CMOS [1], [2]. Recent advances in processing have enabled the fabrication of CNT devices with competitive performance and dimensions. Present challenges to further development of CNFETs call for workflows involving coordinated experimental and simulation efforts. In previous studies, devices compliant with performance targets for sub-5 nm technologies were demonstrated, and an efficient quantum simulation framework capable of describing relevant device physics was developed and demonstrated [3], [4]. Recently, the accuracy of our simulation approach has been validated against experimental measurements of devices using top-gate electrodes to modulate doping in lead extensions [5]. In this work, we explore the impact of using chemical doping on electrical characteristics, and the switching performance limits imposed by SDT in CNFETs with varying temperature, gate length, and CNT diameter.

II. Simulation methodology

A coupled mode-space NEGF solver coupled with a $\vec{k} \cdot \vec{p}$ electronic Hamiltonian is employed to simulate the electrical properties of CNFETs arranged in arbitrary 3D geometries. The $\mathcal{Q}^*$ software package enables computationally efficient simulations of CNFETs including quantum mechanical treatment of elastic and inelastic phonon scattering and electron tunneling phenomena. Carrier injection is described by modeling Schottky contacts to the CNT [6].

The devices studied in this work are based on CNTs with diameters of approximately 1 nm and 2 nm, arranged in a top-gate configuration shown in fig. 1. Boundary conditions applied to electrostatic equations at the contact metal-CNT junctions allow explicitly describing their influence on the electric field throughout the device. Degenerately doped regions along lead extensions are induced either electrostatically, or chemically, to form carrier reservoirs on both sides of the channel. Electrostatic doping is realised by top-gate electrodes placed on top of lead extension regions in order to model structures fabricated in [5]. Devices with chemical doping are modeled using a compensating charge density of $1.17\ \mathrm{nm}^{-1}$ homogeneously distributed along the CNT adjusted to match free carrier densities induced in electrostatically doped devices.

III. Results

Figure 2 shows a comparison of simulated and measured electrical characteristics of long-channel p-type devices based on 2 nm diameter CNTs ($E_G \approx 0.45$ eV), at temperatures of 300K and 10K. Gate-source bias data from simulations has been shifted and adjusted by a multiplicative factor in order to phenomenologically account for non-ideal gate coupling derived from interface traps. We note excellent agreement between experiment and simulation for OFF currents, I_{ON}/I_{OFF} ratio, and the onset of BTBT conduction ($V_{GS} \gtrsim 0$ V).

Figure 3 compares the electrical characteristics of devices using electrostatic and chemical doping at temperatures of (a) 300K and (b) 10K. Devices with chemical doping exhibit larger I_{ON}/I_{OFF} ratios when compared to those electrostatically doped. The fact that chemical doping potential is localised within the CNT results in states that favour larger ON currents and tend to suppress BTBT in OFF states. Figure 3(c) compares the CNT band profile along the device length for an OFF state. Chemical doping induces a steeper profile near source and drain contacts significantly reducing the width of the barrier holes must tunnel across to enter the device near the leads Fermi level, resulting in larger ON currents for the same Schottky barrier height. Conversely, partial screening of the gate action by chemical doping in the channel results in a more gradual band profile on either side of the channel, which suppresses phonon-assisted BTBT that limits OFF current at 300K.

The impact of SDT on the SS at varying gate length, temperature, and CNT diameter is illustrated in fig. 4. Once again we find the case of chemical doping to be advantageous: steeper slopes present in the band profiles near the channel in electrostatically doped devices result in enhanced SDT. SS degradation from electrostatic doping is observed for most gate lengths and temperatures, with largest deviations occurring at shorter gate lengths and higher temperatures (fig. 4(a)-(b)). A lower SS found for devices based on 2 nm diameter CNTs at room temperature is 85 mV/dec for gate lengths of 30 nm or longer. In contrast, fig. 4(c) shows the case of devices based on CNTs with 1 nm diameter ($E_G \approx 0.85$ eV) and chemical doping. Greater electrostatic control is found to be exerted by the gate electrode in these devices, where the long-channel SS reaches the thermal limit of 60 mV/dec at room temperature. Owing to the intrinsically more favourable electronic structure of 1 nm diameter CNTs, SDT is significantly suppressed in these devices, with SS values at room temperature below 65 mV/dec for gate lengths of 10 nm and longer.

IV. Conclusion

CNFETs exhibit excellent electrical characteristics promising suitability for sub-5 nm nodes. However, the full potential of this class of devices has yet to be achieved through refinement of processing techniques and design co-optimization.

Exploration of design parameters and physical analysis enabled by the use of quantum simulations offers insight into device operation at a wide range of temperatures and down to a few Kelvin. We find critical performance metrics of recently reported device designs may be significantly enhanced to achieve larger I_{ON}/I_{OFF} ratios and k_BT/q switching at room temperature for gate lengths of 15 nm or greater. For low-temperature applications, switching may be improved by increasing gate lengths up to around 40 nm, after which the thermal lower limit for SS is reached. Combining advances in processing with the insight offered by physical simulations holds the key for driving CNFETs designs to optimal performance.

References

[1] G. Hills *et al.* "Understanding Energy Efficiency Benefits of Carbon Nanotube Field-Effect Transistors for Digital VLSI," *IEEE Transactions on Nanotechnology*, 2018. DOI: 10.1109/TNANO.2018.2871841

[2] C. Gilardi *et al.* "Extended Scale Length Theory Targeting Low-Dimensional FETs for Carbon Nanotube FET Digital Logic Design-Technology Co-optimization," *IEEE International Electron Devices Meeting*, 2021. DOI: 10.1109/IEDM19574.2021.9720672

[3] G. Pitner *et al.* "Sub-0.5 nm Interfacial Dielectric Enables Superior Electrostatics: 65 mV/dec Top-Gated Carbon Nanotube FETs at 15 nm Gate Length," *IEEE International Electron Devices Meeting*, 2020. DOI: 10.1109/IEDM13553.2020.9371899

[4] A. Sanchez-Soares *et al.* "Top-Gated Carbon Nanotube FETs from Quantum Simulations: Comparison with Experiments," *VLSI-TSA*, 2021. DOI: 10.1109/VLSI-TSA51926.2021.9440104

[5] Q. Lin *et al.* "Bandgap extraction at 10 K to enable leakage control in carbon nanotube MOSFETs," *IEEE Electron Device Letters*, 2022. DOI: 10.1109/LED.2022.3141692

[6] We use the EOLAS proprietary TCAD tool $\mathcal{Q}^*$. https://eolasdesigns.com, Apr 2022.

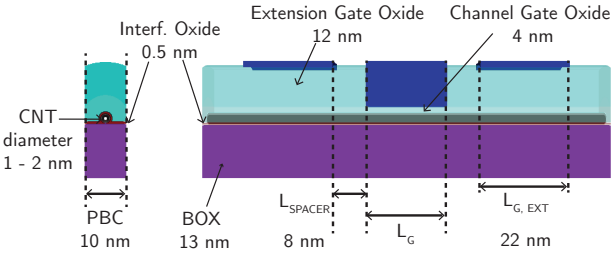


Fig. 1. Schematic of the top-gate device geometry studied. Electrodes located along lead extensions serve to realize electrostatic doping and are omitted when simulating devices with chemical doping. Top oxide is high-κ ($\varepsilon_r = 24$), bottom oxide is SiO_2 ($\varepsilon_r = 3.9$), interf. oxide is Al_2O_3 ($\varepsilon_r = 8$). Schottky contacts to source and drain electrodes are modeled at both ends of the CNT. Periodic boundary conditions (PBC) are employed along the width direction.

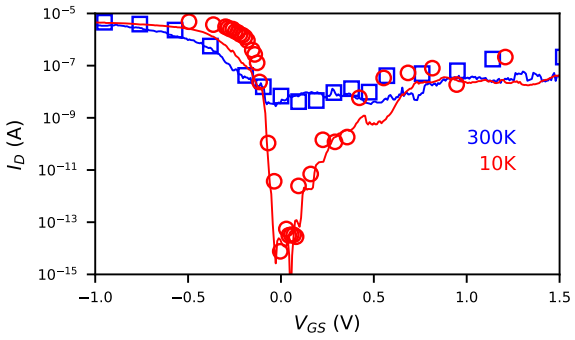


Fig. 2. Current-voltage characteristics of p-type devices at temperatures of 300K (blue) and 10K (red). Drain-source bias is $V_{DS} = 0.3$ V. Full lines correspond to experimental measurements, and symbols correspond to simulations.

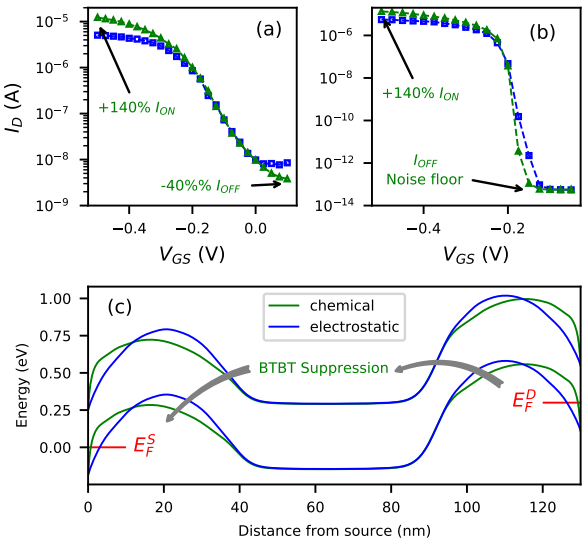


Fig. 3. Current-voltage characteristics of p-type devices at temperatures of (a) 300K and (b) 10K for designs with either chemical (green) or electrostatic (blue) doping. (c) Band edges along the length of devices for an OFF state. Source Fermi level (E_F^S) is at 0 eV, and drain Fermi level (E_F^D) at 0.3 eV.

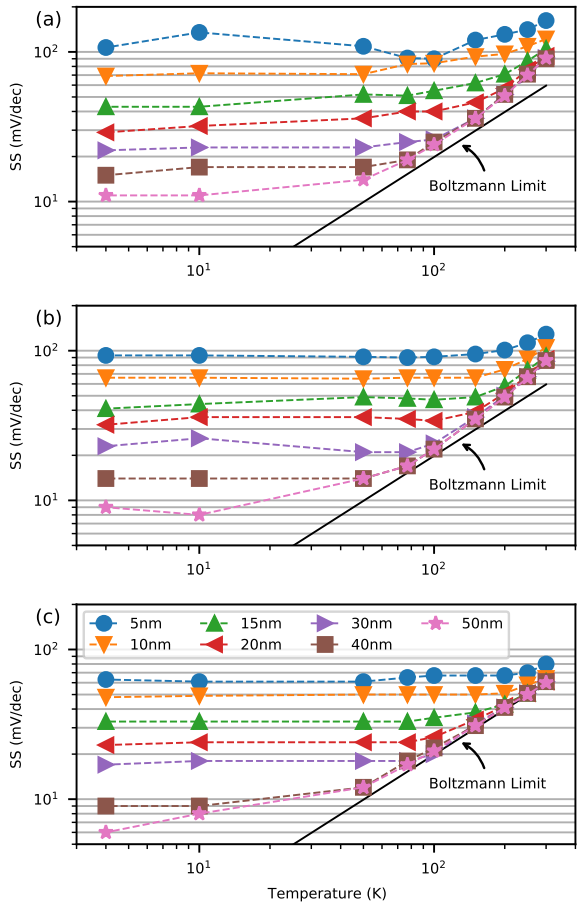


Fig. 4. Subthreshold swing (SS) versus temperature for devices based on 2 nm diameter CNTs with (a) electrostatic doping or (b) chemical doping, and (c) 1 nm diameter CNTs and chemical doping. Lengths reported in legend correspond to channel length.

28nm FDSOI MEOL Parasitic Capacitance Segmentation using Electrical Testing and Semiconductor Process Modeling

B. Vianne, B. Guillo-Lohan, V. Quenette, B. Legoix, B. Vincent*
STMicroelectronics, Crolles, France
*Coventor, a Lam Research Company, Villebon sur Yvette, France
email: benjamin.vianne@st.com

Abstract

This paper describes an extraction methodology for segmenting the different contributions to interconnect and contact driven parasitic capacitance present on a 28nm Fully Depleted Silicon On Insulator technology. The segmentation was enabled by creating specific test structures that had been designed, processed, and electrically tested across full wafer mappings. A 3D semiconductor process model, including capacitance extraction, was subsequently built and calibrated using the statistical distribution of actual silicon data. Once fully calibrated (< 3% mean deviation to actual data), the model was used to understand the sensitivity of parasitic capacitance to specific process/design parameters and to enable Design Technology Co-Optimization.

Introduction

As MOS transistor dimensions continue to decrease, parasitic capacitance introduced by source/drain contacts are increasing [1] and need to be accurately estimated during logic circuit design. Analytical [2-3], 2-D [4] or 3-D TCAD models [5] are generally used to estimate the fringe capacitances in MOSFET devices, and to understand their relationship with process parameters such as gate height or source/drain epitaxy thickness. However, in these models, assumptions are made to simplify the geometry of devices compared to the actual morphology on silicon. In this paper, we propose to quantify the parasitic capacitance between source/drain contacts and gates on a 28nm FDSOI MOSFET structure and will do so by using a geometrically accurate 3-D virtual process model and calibrating it against actual wafer measurements.

Test structure details and electrical-test segmentation

Six transistor designs were considered with different configurations of channel widths (W) and poly to contact distances (Po2Co). In all structures, the transistors are isolated from the bulk substrate to decorrelate the MOS contribution during parasitic extraction. Each transistor design is duplicated in two configurations:

-Configuration A, with poly on shallow trench isolation (STI) and source/drain contacts on STI;

-Configuration B, with poly on STI but without source/drain contacts.

Minimum, median and maximum capacitance values from the Gate to the Source/Drain were extracted from a silicon-based e-test performed on 98 dies on a wafer. The results are reported in Figure 1 for the two de-embedded test structures. The percentage contribution of the contact capacitance as a portion of the total parasitic capacitance is calculated using the equation $[(C_A - C_B)/C_A]$. Figure 1 illustrates that 10-30% of the total capacitance is based on the contact contribution, depending on the design considered.

Virtual Process Model vs Measured Capacitances

SEMulator3D® virtual semiconductor process models were built in order to better understand the correlation between process parameter changes and the effect on various parasitic capacitances. Three different SEMulator3D® models were considered, as depicted in Figure 2. In addition, a standard parasitics extraction procedure (PEX) was used as a reference. For each of the structures simulated, parasitic capacitance between the gate and the source interconnects and the gate and the drain interconnects were extracted using SEMulator3D®.

For each transistor design and each of the three model types detailed above, Figure 3 reports the simulated Gate to Source/Drain parasitic capacitance compared to the normalized (across multiple transistors) mean Gate to Source/Drain Capacitance extracted from Si wafer measurements. The extruded TCAD geometric simulation model shows a similar mean deviation as PEX (12-14%) when compared to silicon data. The more realistic SEMulator3D silicon 3D model shows a much better correlation with the actual silicon data, with the error difference not exceeding 7%. This demonstrates the importance of properly matching the geometry of the silicon structure to the structure used in simulation, in order to accurately extract electrical (parasitic capacitance) data during simulation.

Model Calibration and Process Sensitivity Analysis

The silicon realistic 3D model was further improved by completing an additional automatic model calibration step using silicon e-test data. A calibrated model ensures that the model will more accurately reflect process behavior seen in silicon.

Nine (9) different process parameters were then selected for further analysis, as detailed in Figure 4.

A large Virtual Design Of Experiment was executed by completing 200 virtual experiments on each of the 12 devices. In the experiment, parameter values were varied by generating and assigning Monte Carlo normal distribution values to each of the nine selected process parameters.

Silicon wafer mean values for all 12 devices were collected and compared to the simulation data. A regression analysis was performed to calibrate the 9 process parameters and further optimize the correlation to Si data. Using the calibrated results (Figure 3), a “Calibrated Silicon Realistic Model” was simulated and produced a deviation to silicon e-test data of only 2.9 and 5.9% (mean and max values, respectively). This deviation is lower than any of the other models.

The statistical data obtained from the Design of Experiments (a total of 2400 virtual experiments) was subsequently used to determine the sensitivity of the calculated parasitic capacitance values to each of the selected process parameters. Focusing on a real MOS structure (in-

cluding interconnects, contacts and active components), Table 1 highlights the weight, from highest to lowest, of the most significant process parameters and their cross-terms that affect the parasitic contribution of Gate to Source/Drain capacitance. The table highlights the most important parameters that can be controlled or adjusted to reduce parasitics.

Conclusions

Independent parasitic contributions of contacts and interconnects to Gate to Source/Drain capacitance were identified, using test structures specifically fabricated, tested and simulated for that purpose. The simulation results included a large statistical data set (2400 experiments). This data set was used to build the correlation between 9 different process parameters and measured capacitance on various test structures (including a completed MOS structure). Seven significant process parameters (including cross term factors) were identified as the most important parameters to control, in order to reduce and better control the fringe capacitance of MOSFETs. The simulation results were highly predictive (<3% mean deviation) when compared to actual wafer data, as long as accurate, well-calibrated structural geometries were used during the process model simulation and electrical analysis.

Acknowledgment -The authors would like to thank H. El Dirani and J.-P. Carrère from STMicroelectronics Crolles TDP for the technology and design support.

References

[1] P. K. Chatterjee et al, "The impact of scaling laws on the choice of n-channel or p-channel for MOS VLSI," in IEEE Electron Device Letters, vol. 1, no. 10, pp. 220-223, Oct. 1980, doi: 10.1109/EDL.1980.25295.

[2] N.R. Mohapatra, M.P. Desai, S.G. Narendra, and V. Ramgopal Rao, "Modeling of Parasitic Capacitances in Deep Submicrometer Conventional and High-K Dielectric MOS Transistors", IEEE Trans Electron Devices, vol. 50, no. 4, pp. 959-966, Apr. 2003, doi: 10.1109/TED.2003.811387.

[3] M.J. Kumar, V. Venkataraman, and S.K. Gupta, "On the parasitic gate capacitance of small-geometry MOSFETs", IEEE Trans Electron Devices, vol. 52, no. 7, pp. 1676-1677, Jul. 2005, doi: 10.1109/TED.2005.850630

[4] B. Ricco, R. Versari and D. Esseni, "A novel method to characterize parasitic capacitances in MOSFET's," in IEEE Electron Device Letters, vol. 16, no. 11, pp. 485-487, Nov. 1995, doi: 10.1109/55.468275.

[5] C. R. Manoj, Angada B. Sachid, Feng Yuan, Chang-Yun Chang, and V. Ramgopal Rao, "Impact of Fringe Capacitance on the Performance of Nanoscale FinFETs", IEEE Electron Device Letter, vol. 31, no. 1, pp. 83-85, Jan. 2010, doi: 10.1109/LED.2009.2035934.

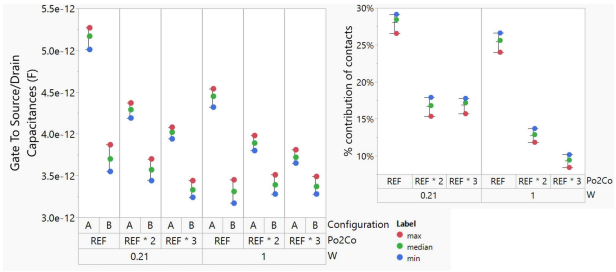


Figure 1: Capacitance values extracted from de-embedded structures and contact contribution (in %) of the parasitic capacitances.

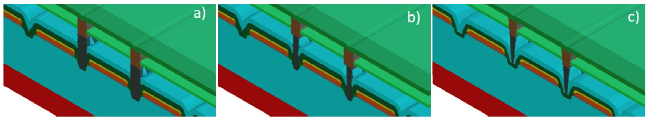


Figure 2: Three types of process models: a) 3D extruded geometric model from the 2D layout, b) 3D extruded geometric model from the 2D layout with Critical Dimension correction matching of the top CDs for the Via/Lines XTEMs, c) 3D silicon realistic model that includes realistic etch models/profiles (calibrated to XTEM), lithography corrections (CD) and emulation (corner rounding).

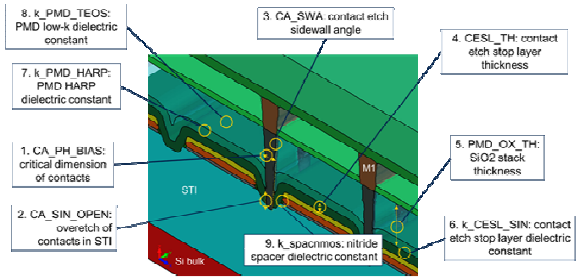


Figure 4: Process parameters from the silicon realistic model that were used for further calibration.

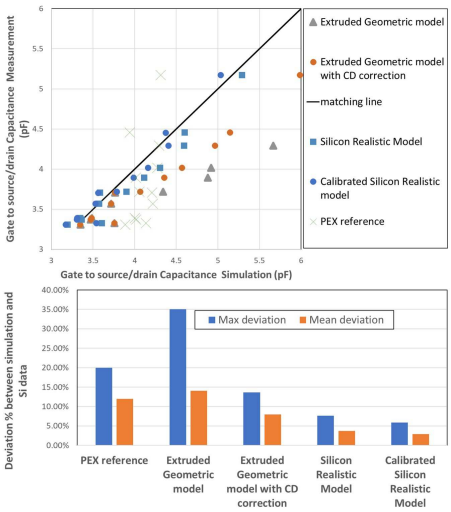


Figure 3: Simulated parasitic capacitances on 12 different devices for different models (top), and deviation % between simulation and Si data values (bottom).

Parameter	P-Value	Weight
Contact Etch Stop Layer Dielectic Constant	0.0128	-1.098
Contact Via sidewall angle	<0.001	-0.3113
(Contact Etch Stop Layer Dielectic Constant) * (PMD HARP dielectic constant)	0.00246	0.167
(Contact Etch Stop Layer Dielectic Constant) * (PMD low-k dielectic constant)	0.003717	0.1583
Contact Via CD	<0.001	0.1155
(Contact Etch Stop Layer Thickness) * (PMD low-k dielectic constant)	<0.001	0.006642
(PMD HARP dielectic constant) * (PMD low-k dielectic constant)	<0.001	0.006642

Table 1: Process parameters and cross-term parameters ranked based upon their weight in affecting Gate to Source/Drain parasitic capacitance.

Simulation-based study on characteristics of dual vertical transfer gates in sub-micron pixels for CMOS image sensors

Wook Lee[1], Seonghoon Ko[1], Uihui Kwon[1], HyunChul Kim[2], Dae Sin Kim[1]
[1] Computational Science and Engineering Team, Innovation Center
[2] Semiconductor R&D Center
Device Solution Business, Samsung Electronics Co.,
Hwasung-si, Gyeonggi-do, Republic of Korea.
* Email Address: wook6.lee@samsung.com.

Abstract— **Recently dual vertical transfer gates (VTGs), used in sub-micron pixels with full-depth deep-trench isolation (FDTI), have demonstrated superior performance in CMOS image sensors such as improvement of full well capacity (FWC) and charge transfer, as compared to a single VTG. In this work, we investigate characteristics of both pixel schemes based on two design examples, which is carried out using extensive 3D TCAD simulation and automated multi-objective optimization flow with various photodiode implantation conditions satisfying certain design specifications. Simulation results reveal that dual VTGs better control electrostatic potentials along the charge transfer path like a 3D fin-shaped transistor. The enhanced gate controllability also makes the VTG off potential insensitive to the nearby doping concentrations, which is not the case for the single VTG pixel, and thus provides more room for boosting FWC in the photodiode design according to the Pareto front analysis.**

Keywords— TCAD Simulation, CMOS Image Sensor, Vertical Transfer Gate, Photodiode, Full Well Capacity, Charge Transfer.

INTRODUCTION

Among various sub-micron pixel architectures in CMOS image sensors, the unit cell that is composed of the single VTG with deep photodiode and separated by FDTI structure has gained increasing popularity due to many favorable sensor characteristics, such as high FWC and low crosstalk [1-3]. Recently the dual-VTG scheme has been successfully applied to a 0.6μm pitch pixel for the first time, and achieved both FWC increase by 60% and improved image lag as compared to the single VTG [4]. In this work, we analyze the physical origin of these performance gains by comparing each pixel design based on TCAD simulation.

Due to three dimensional nature of charge transfer from the deep photodiode to the shallow floating diffusion (FD) node via the VTG, optimization of these device components in terms of the pixel layout and process conditions are extremely challenging, not to mention other in-pixel transistors affecting all each other. Elaborate engineering of doping profile in the photodiode and its adjacent region is required such that signal electrons should not see any potential barriers along the transfer path [5]. The deep photodiode itself is formed by a sequence of n-type implant with different energies up to a few MeV to ensure high FWC and so its potential profile also needs to be well tuned not to leave any residual electrons in it.

Given each pixel layout based on single and dual VTGs given in Fig. 1 where the photodiode is formed deep in Silicon per pixel and separated by FDTI, we optimize the pixel with three design specifications; (1) FWC needs to be maximized, or at least be larger than some threshold, e.g., >5,000e-, (2) no residual electrons be sensed in the readout operation, (3) the potential hump on the transfer path, if any, be smaller than 100mV when the VTG is on. Several implantation conditions in the photodiode and around the VTG region are varied as input parameters, and optimization is performed using our in-house automated machine-learning based optimization framework with extensive 3D TCAD simulation to speed up the pixel design and explore the design space effectively [6].

RESULTS & DISCUSSION

The transfer path of electrons moving from the deep photodiode center to the FD node is traced based on spatial potential gradients, and its potential variations as a function of Silicon depth is plotted in Fig. 2 for each pixel under the same implant conditions. In the dual-VTG case, electrons transfer through the channel between VTGs, and the channel potential is better controlled by the gate voltage than the nearby doping profile. Electrical characteristics are also compared in Table. 1. Not only does FWC increase by 3,000e- mainly due to different TGLSO levels, but also the potential around the VTG region varies more smoothly leads to the better transfer capability in the dual-VTG pixel.

Fig. 3 shows full optimization results of the single-VTG pixel. Each data point represents one solution with different optimal implant conditions satisfying the aforementioned target specifications. PDMAX and TGLSO is the highest potential in the photodiode and the VTG off potential, respectively, and the difference between these two is roughly proportional to FWC. In Fig. 3 (a), it is important to note that the maximal FWC, which can be achieved without the transfer problem, is basically limited to some extent in the single-VTG pixel, regardless of combinations of the doping profile in the photodiode. In other words, if one attempts to increase the net doping concentrations of the photodiode to increase FWC beyond the Pareto front, the transfer characteristics are likely to be degraded out of specification, or since TGLSO also increases proportionally with PDMAX as presented in Fig. 3 (b), it eventually results in no much gain in FWC.

Interestingly, this tradeoff can be greatly alleviated in the case of dual VTGs as shown in Fig. 4. Due to the enhanced gate controllability in dual VTGs like a three dimensional fin-type transistor, TGLSO does not vary much around a potential of -0.35V even with diverse input implant conditions as shown in Fig. 4 (b), while PDMAX is still dominated by the photodiode net-doping. Therefore, one can design the dual-VTG pixel to have a higher FWC without sacrificing the charge transfer, as illustrated by the Pareto front in Fig. 4 (a).

CONCLUSION

In summary, we conduct comparative study on electrical characteristics of single- and dual-VTG based sub-micron pixels using 3D TCAD simulation and automated multi-objective optimization flow. The Pareto front analysis based on massive simulation data shows that dual VTGs can greatly improve the sensor characteristics in terms of both FWC and charge transfer because of the enhanced gate controllability, by reducing the conventional performance tradeoff that presents in the single-VTG pixel.

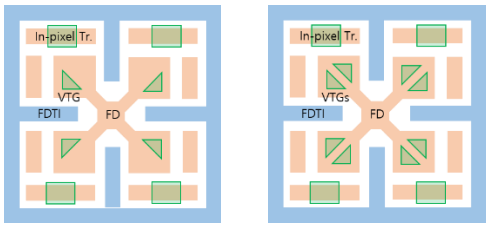


Fig. 1. Schematics of 2x2 pixel layout based on (a) single VTG and (b) dual VTGs.

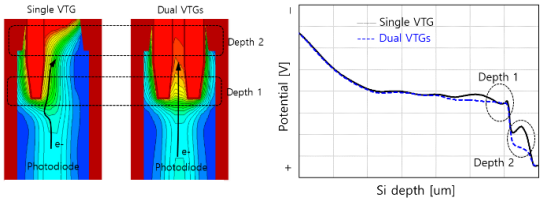


Fig. 2. Potential profile and electron transfer path in single- and dual-VTG pixels.

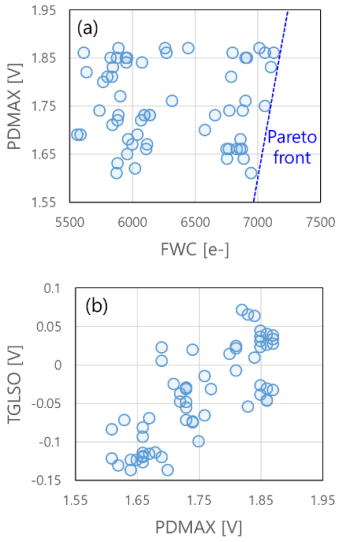


Fig. 3. Pareto front analysis of single-VTG pixel: (a) FWC vs. PDMAX (b) PDMAX vs. TGLSO.

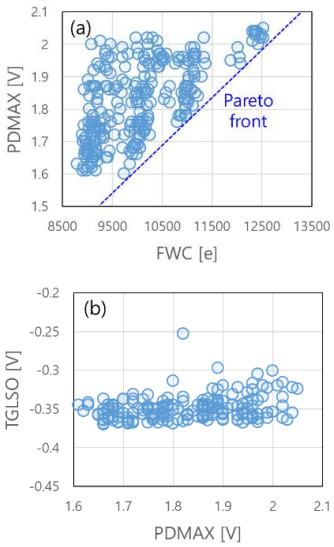


Fig. 4. Pareto front analysis of dual-VTG pixel: (a) FWC vs. PDMAX (b) PDMAX vs. TGLSO.

Table 1. Simulated electrical characteristics of pixel design examples given in Fig. 2.

Type	Single-VTG pixel	Dual-VTG pixel
FWC [e-]	6250	9202
Residual electrons [e-]	0	0
TGLSO [V]	-0.04	-0.35
PDMAX [V]	1.79	1.78
Max. Potential hump in Depth 1 [mV]	61	0
Max. potential hump in Depth 2 [mV]	155	0

REFERENCES

[1] Y. Kim, *et al.*, "A 1/2.8-inch 24Mpixel CMOS image sensor with 0.9μm unit pixels separated by full-depth deep-trench isolation," *IEEE International Solid - State Circuits Conference*, 2018.

[2] D. Park, *et al.*, "A 0.8 μm Smart Dual Conversion Gain Pixel for 64 Megapixels CMOS Image Sensor with 12k e- Full-Well Capacitance and Low Dark Noise," *IEEE International Electron Devices Meeting*, 2019.

[3] J. Park, *et al.*, "1/2.74-inch 32Mpixel-Prototype CMOS Image Sensor with 0.64μm Unit Pixels Separated by Full-Depth Deep-Trench Isolation," *IEEE International Solid-State Circuits Conference*, 2021.

[4] J. Yun, *et al.*, "A 0.6 μm Small Pixel for High Resolution CMOS Image Sensor with Full Well Capacity of 10,000e- by Dual Vertical Transfer Gate Technology," *IEEE Symposium on VLSI Technology and Circuits*, 2022.

[5] S. Kim, *et al.*, "Potential Engineering to Enhance Transfer Characteristics of Advanced CIS Pixel based on VTG - FDTI scheme," *International Conference on Simulation of Semiconductor Processes and Devices*, 2021.

[6] J. Yoo, *et al.*, "Machine-Learning based TCAD Optimization Method for Next Generation BCD Process Development," *International Symposium on Power Semiconductor Devices and ICs*, 2021.

Compact Model of a Metal Oxide Molecule Sensor for Self-Heating Control

Yohsuke Shiiki
Electronics and Electrical Engineering
Keio University
Yokohama, Japan
shiiki@iskr.elec.keio.ac.jp

Shintaro Nagata
Department of Applied Chemistry
The University of Tokyo
Tokyo, Japan

Tsunaki Takahashi
Department of Applied Chemistry
The University of Tokyo
Tokyo, Japan

Takeshi Yanagida
Department of Applied Chemistry.
The University of Tokyo
Tokyo, Japan

Hiroki Ishikuro
Electronics and Electrical Engineering
Keio University
Yokohama, Japan

***Abstract*—Molecule sensors made from metal oxide have notable advantages of low cost and small size, which are suitable for applications accumulating a large amount of data represented by IoT. In order to collect data from wide regions at any time, establishing a low-power sensing system is inevitable. However, most metal oxide sensors require high sensing temperature to cause chemical reactions with surrounding molecules, so that an external heater is integrated with those sensors and around 1 W or higher power is consumed. One candidate solution to decrease power consumption is using a self-heated sensor. Since the heating area is limited in the sensor itself, mW order sensing operation can be achieved. Besides, it is also possible that the target molecule is selected by changing sensor temperature. The self-heated sensor needs careful temperature control, and it should be handled by an interface analog circuit. Therefore, its compact model in MATLAB or Verilog-A should be created to minimize total power consumption and verify its sensitivity to interface circuit performance. In this paper, a developing self-heated resistive molecule sensor was shown and a formula to simulate sensor resistance was established. The resistances obtained with measurement probe in experiments and simulated values were compared.**

I. Introduction

Compared with most conventional semiconductor molecule sensors, self-heated molecule sensors are beneficial for developing low-power system because they consume small power to increase sensing temperature [1]. In addition, there have been some reports that the sensor configured to different temperature reacts to different molecules [2]-[4]. A self-heated sensor only heats a limited sensing area, thus short time constant for controlling temperature can be realized. If the temperature of those sensors is controlled precisely, the number of detectable molecules will increase. One of the obstacles to realizing the system is to control sensor temperature by feedback power from the interface circuit. The main operation of the circuit is keeping the sensitivity and controlling temperature by feedback voltage or current while consuming a small amount of power. Designing the dedicated circuit for the self-heated sensor, and making sensors compact models for circuit simulation is inevitable. Moreover, compact modeling is used for analyzing sensor data in advance by integrating many sensors into a simulation environment (Fig. 1).

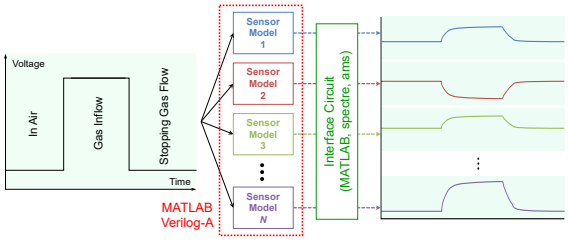
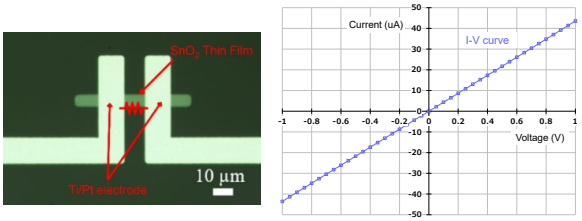


Fig. 1. Simulated sensor responses.



(a) (b)

Fig. 2. (a) Microscope image of the self-heated sensor device. (b) I – V curve of the device

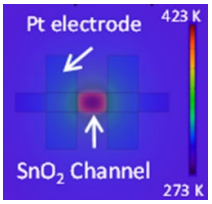


Fig. 3. Thermal simulation results.

II. Developed Sensor

In Fig. 2, the developed self-heated sensor and its measured I-V characteristics are shown. The I-V curves indicates the sensor is an ohmic device. SnO_2 thin film (20nm) is used as the sensing material, and an electrode is made of Ti/Pt. The resistance of SnO_2 film is changed if it adsorbs certain molecules. The narrow sensor shape and 20nm thinness makes it operate as a self-heated sensor. To verify the self-heated characteristics, thermal simulation was carried out with COMSOL Multiphysics software, and it was verified that only the sensor spot became high temperature as shown in Fig. 3.

III. Compact Modeling

The compact model of the self-heated sensor is shown in Fig. 4. While the sensor's temperature was controlled with an external heater in the previous works [5], the feedback loop to cause temperature fluctuation derived from its own resistance

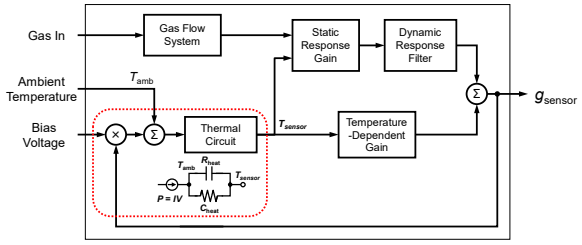


Fig. 4. Self-heated molecule Sensor model.

change was added. To simplify the model, the number of input gas was limited to one. The gas flow system and the dynamic response filter are low-pass filters. In this work, the gas flow system was removed because the dynamic response filter was dominated filter. The static response gain and temperature-dependent gain are represented as follows:

$$G_{gas} = k_{1T} \cdot e^{-E_{A1}/kT} \cdot C^{n_1 kT} \quad (1)$$

$$G_{base} = G_{0T} \cdot e^{-E_{A0}/kT} \quad (2)$$

where k and T represent Boltzmann constant and absolute temperature. E_{A0} and E_{A1} are activation energies of the baseline conductance and of the change conductance by a target gas. k_{1T}, G_{0T}, and n_1 are coefficients. C represents the concentration of input gas. The thermal circuit consists of the thermal resistance and the thermal capacitance. Since it is quite difficult to measure the thermal circuit parameters, simulation results obtained from COMSOL Multiphysics software were substituted. The input voltage representing 1ppm gas concentration corresponds to 1V and the voltage follows linearly to the gas concentration.

IV. Experiment and Simulation Results

The unknown model parameters: E_{A0}, E_{A1}, k_{1T}, G_{0T}, and n_1 were determined by fitting them to experimental data with equations (1), (2). In the experiment, the target gas was NO_2, and the surrounding gas was N_2. In the experiment, resistance was measured under different temperatures of 150 °C, 200 °C, and 250 °C. Since it was difficult to monitor the sensor's surface temperature accurately, an external heater was used. Then, the bias voltage was set to 1V so as not to cause self-heating. In each temperature condition, gas concentration was set to 0ppm, 10ppm, 20ppm, 50ppm, and 100 ppm, and a steady-state was measured. TABLE I shows adjusted parameters. In Fig. 5, the comparison between experimental data and calculated data from the determined parameters is shown.

With determined parameters in the above experiments, the dynamic responses were compared. The ambient the temperature was set to 100 °C and the sensor was biased with

TABLE Ⅰ. Parameters determined by experiments and COMSOL Multiphysics software.

Parameters	Value
G_{0T}	0.0001095 [S]
k_{1T}	-0.003043 [S/ppm]
E_{A0}	0.01567 [eV]
E_{A1}	0.2334 [eV]
n_1	3.176 [K^{-1}]
R_{heat}	29.7 [K/mW]
C_{heat}	0.0297 [uJ/K]

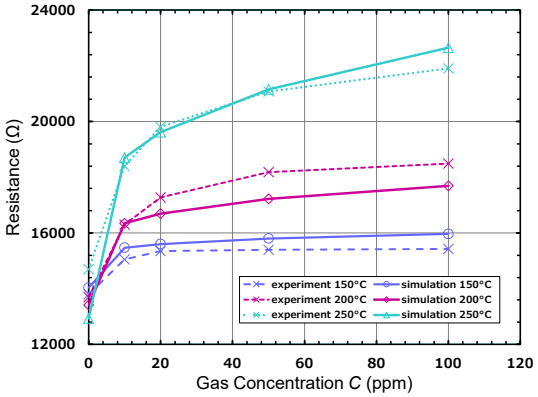


Fig. 5. Comparison of static response between the simulated values with the fitting parameters and experiment data.

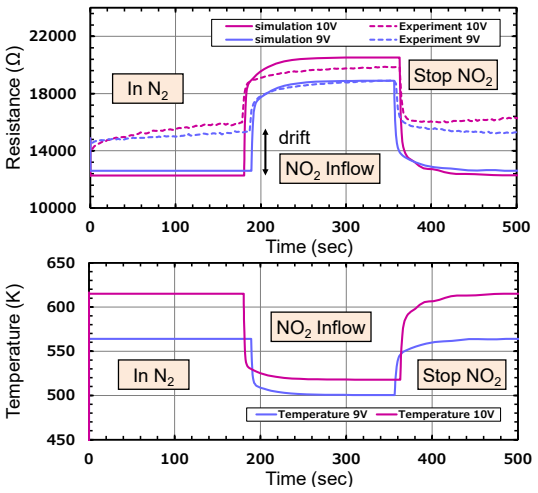


Fig. 6. Comparison of dynamic responses between the simulated values with the fitting parameters and experiment data.

9V and 10V. Fig. 6 shows the resistance comparison between simulation and experiment values. Although the sensor drift effects were seen in the no gas phase, the static sensor response value in the simulation followed the experiment resistance change. It was also shown that sensing temperature in the simulations was fluctuate as high as 90°C when the sensor reacted to NO_2 under a constant bias voltage condition.

References

[1] T. Tanaka, K. Tabuchi, K. Tatehora, Y. Shiiki, S. Nakagawa, T. Takahashi, R. Shimizu, H. Ishikuro, T. Kuroda, T. Yanagida, K. Uchida, "Low-Power nad ppm-Level Multimolecule Detection by Integration of Self-Heated Metal Nanosheet Sensors," *IEEE Trans. Electron Devices*, vol. 66, no. 12, pp. 5393-5398, Dec. 2019.

[2] H. Liu, G. Meng, Z. Deng, K. Nagashima, S. Wang, T. Dai, L. Li, T. Yanagida, X. Fang, "Discriminating BTX Molecules by the Nonselective Metal Oxide Sensor-Based Smart Sensing System," *ACS Sensors* 2021, 6, 4167−4175.

[3] H. Liu, Y. He, K. Nagashima, G. Meng, T. Dai, B. Tong, Z. Deng, S. Wang, N. Zhu, T. Yanagida, X. Fang, "Discrimination of VOCs molecules via extracting concealed features from a temperature-modulated p-type NiO sensor," Sens. Actuators B 293 (2019), 342-349.

[4] F. Hossein-Babaei, A. Amini, "A breakthrough in gas diagnosis with a temperature-modulated generic metal oxide gas sensor," Sens. Actuators B 166-167 (2012) 419-425.

[5] E. Llobet, X. Vilanova, J. Brezmes, D. López, X. Correig, "Electrical equivalent models of semiconductor gas sensors using PSpice," Sens. Actuators B 77 (2001) 275-280

Deriving a novel methodology for Nano-BioFETs and analyzing the effect of high-k oxides on the amino-acids sensing application

Rakshita Dhar[1], Naveen Kumar[1], Cesar Pascual Garcia[2], Vihar Georgiev[1]

[1]Device Modelling Group, James Watt School of Engineering, University of Glasgow, UK

[2]Nano-Enabled Medicine and Cosmetics group, Materials Research and Technology Department, Luxembourg Institute of Science and Technology (LIST), Belvaux, Luxembourg

*Correspondence: r.dhar.1@research.gla.ac.uk, naveen.kumar@glasgow.ac.uk, cesar.pascual@list.lu, vihar.georgiev@glasgow.ac.uk

***Abstract*— In this paper, a novel methodology is presented with the analytical simulations of BioFETs using the Gouy-Chapman-Stern and Modified Site-binding model. The derived approach is used to detect different amino acids such as Arginine (R), Aspartic Acid (D) and Proline (P), functionalized with the help of a linker over the gate-oxide. The performance of the BioFETs is optimized while analyzing the effect of high-k dielectrics as the gate oxide. High-k oxides are responsible for tuning the parameters such as sensitivity, surface potential and intrinsic buffer capacity. The variation of differential capacitance with the second gradient of drain current and surface potential are used to identify the signatures of different amino acids. The proposed method can be helpful in defining an efficient method for protein sequencing.**

Keywords- ISFET, nano-biosensing, PMI, TCAD simulation.

I. INTRODUCTION

Since 1952, when Shockley invented Field-Effect Transistors (FETs); there have been undergoing many significant modifications to cater to the needs of sensing applications. In 1970, Bergveld developed an ion-sensitive Field-Effect Transistor (ISFET) which is used for ion detection in a chemical environment. With further development of technology, novel functionalizations for ISFET devices have been fabricated for protein and DNA detection. Various simulations have been carried out to further deepen the knowledge of the impact of functionalizations on the performance parameters [1]. In this paper, analytical simulations have been carried out to effectively simulate the Gouy-Chapman-Stern model with the Site-Binding model for different amino acid functionalized on various types of gate oxides. Further, we plan on integrating numerical TCAD simulations with this work.

II. METHODOLOGY

The analytical simulations are used to calculate the zeta potential (Ψ_ξ), sensitivity factor (α) and intrinsic buffer capacity (β) values for all gate oxides [2]. The following methodology is used for a particular gate oxide with or without functionalized with the amino acids. First, the surface charge density is calculated using the Gouy-Chapman-Stern model and Boltzmann-Poisson model (σ_{DL}).

$$\sigma_{DL1} = qN_S\left(\frac{cH_s^2 - K_aK_b}{K_aK_b + K_bcH_S + cH_S^2}\right)$$

$$\sigma_{DL2} = qN_S\left(\frac{-cH_S K_a - K_aK_b}{K_aK_b + K_bcH_S + cH_S^2}\right)$$

$$\sigma_{DL3} = qN_S\left(\frac{-K_a}{K_a + cH_S}\right)$$

Where, $cH_s = cH_B exp\frac{-\Psi_0}{2V_T}$, $cH_B = 10^{-pH_B}$, pH_B is the bulk pH, Ψ_o is the surface potential, N_S is the total surface states, K_a & K_b are the dissociation constants of the corresponding reactive sites and V_T is the thermal voltage. The σ_{DL} is equated and calculated iteratively with the σ_o from the site-binding model. Bi-section method is adopted to find pH_{pzc} and Ψ_ξ for a particular surface charge density following the below-mentioned relation.

$$\Psi_0 = \Psi_{Stern}+\Psi_\xi = \frac{Q_0 \sinh(\Psi_\xi/V_T)}{C_{stern}}+\Psi_\xi$$

Where Ψ_{Stern} is the potential drop across the stern layer and C_{Stern} is the capacitance of the stern layer. Once Ψ_0 is obtained then α, β, and $\frac{\partial^2\Psi_0}{\partial pH^2}$ and can be calculated as shown in Ref. [1][2]. A high-aspect-ratio FinFET is used to calculate the depletion width and drain current variation while solving the continuity equation across the electrolyte-oxide-semiconductor interfaces using surface potential as the variable parameter [2].

III. RESULTS AND DISCUSSIONS

Fig. 1 shows a schematic of BioFET with the oxide region and the functionalized amino acid (Carboxyl-terminal Immobilized Aspartic Acid). SiO_2, HfO_2 and TiO_2 have been selected as the gate oxides to analyze the effect of high-k dielectric on the sensitivity of the sensor. From Fig. 2(a), point-of-

zero-charge (pH_{pzc}) for SiO_2, HfO_2 and TiO_2 are clearly understood as the pH at which the zeta potential crosses the zero value. As shown in Fig. 2 (b), α is minimum at pH_{pzc} but for HfO_2 it's around 0.6 (higher than SiO_2 and TiO_2) since the change in Ψ_ξ w.r.t. pH is higher for enhanced sensitivity. Also, the increment in α is observed for all gate oxides before and after pH_{pzc}. β has a similar trend w.r.t. pH as it denotes the change in σ_0 w.r.t. pH_S [Fig. 2(c)]. Fig. 3(a) shows the Ψ_0 variation for all three gate oxides considered which follows the Ψ_ξ with the drop across the stern layer. Fig. 3(b) is the 2nd order gradient of Ψ_o that shows the transition of reactive sites over the oxide from protonation to deprotonation clarifying the pH_{pzc} value. Fig. 3(c) shows the total capacitance (C_T) which consists of stern capacitance, diffuse-layer capacitance and intrinsic oxide capacitance. TiO_2 shows better resolution for ion-sensing with higher capacitance and non-linearity of Ψ_o due to the large difference between the affinity constants. Considering the full coverage, Fig. 4 represents the depletion width, drain current (I_{SD}) and 2nd order gradient of I_{SD} $\left(\frac{\partial^2 I_{SD}}{\partial pH^2}\right)$ for SiO_2 which varies depending on the affinities of Carboxyl/Amine-term. immobilized Aspartic Acid (D) [σ_{DL1}], Arginine (R) [σ_{DL2}], and Proline (P) [σ_{DL3}]. Different site-binding relations are used to calculate the Ψ_o. The affinity of the remaining amine/carboxyl sites and sidechains are responsible for the distinct fingerprints for the amino acids in the form of I_{SD} and $\frac{\partial^2 I_{SD}}{\partial pH^2}$. Such a simplified model can enhance the possibility of using FET-based sensors for model proteomics.

REFERENCES

[1] Medina-Bailon, C.; Kumar, N.; Dhar, R.P.S.; Todorova, I.; Lenoble, D.; Georgiev, V.P.; García, C.P. Comprehensive Analytical Modelling of an Absolute pH Sensor. *Sensors* 2021, *21*, 5190.

[2] R.E.G. van Hal, J.C.T. Eijkel, P. Bergveld, A general model to describe the electrostatic potential at electrolyte oxide interfaces, Advances in Colloid and Interface Science, Volume 69, Issues 1–3, 1996, Pages 31-62.

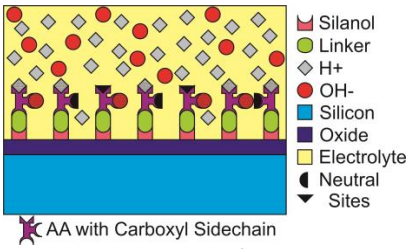


Fig.1. Schematic diagram of the BioFET sensor with functionlized amino acid over the Oxide

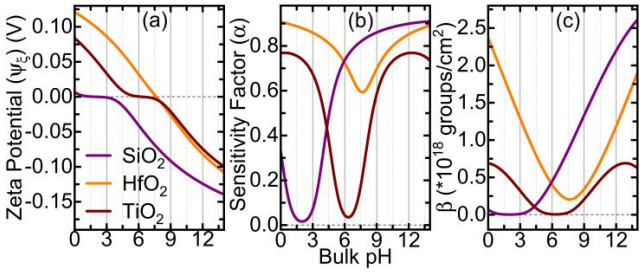


Fig.2. (a) Zeta Potential (Ψ_ξ), (b) Sensitivity factor (α) and (c) intrinsic buffer capacity (β) variation with respect to pH for different gate oxides

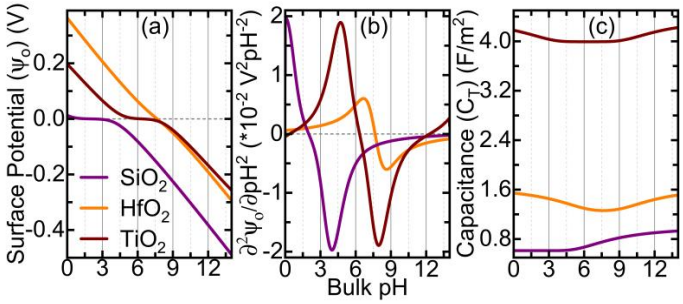


Fig.3. (a) Surface Potential (Ψ_0) (b) 2nd order gradient of Ψ_0 $\frac{\partial^2 \Psi_0}{\partial pH^2}$ and (c) Total capacitance (C_T) Vs pH graphs for different gate oxides

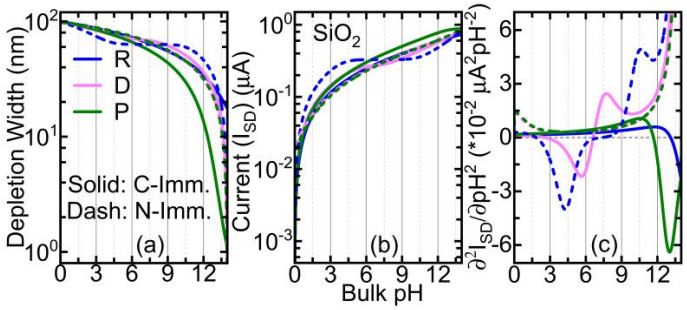


Fig.4. (a) Depletion width (b) drain current (I_{SD}) and (c) 2nd order gradient of I_{SD} $\left(\frac{\partial^2 I_{SD}}{\partial pH^2}\right)$ Vs pH graphs for Arginine (R), Aspartic Acid (D) & Proline(P) amino acids having SiO_2 as gate oxide with Carboxyl group immobilized (C-Imm) and amine group immobilized (N-Imm)

Ab initio modeling of photodetectors based on van der Waals heterostructures

Jiang Cao, Sara Fiore, Cedric Klinkert, Mathieu Luisier
Integrated Systems Laboratory, *ETH Zürich* , 8092 Zürich, Switzerland
jiang.cao@iis.ee.ethz.ch, mluisier@iis.ee.ethz.ch

Introduction: Strong light-matter interactions in van der Waals heterostructures (vdWHs) made of two-dimensional (2-D) materials, especially transition metal dichalcogenides (TMDCs) [1], have brought new perspectives to the research on optoelectronic devices. Stacking different TMDC monolayers into vdWHs give rise to type-II band alignments with (quasi-)direct-gap [2]. This facilitates the optical generation of interlayer excitons, where electrons and holes separate on ultra-short time scales and locate on different layers with extremely long lifetime [3]. Lower electric fields are needed to separate interlayer electron-hole (e-h) pairs with respect to intralayer ones due to the reduced Coulomb interaction [2]. These advantages open the door for photodetectors with high responsivities at lower bias. In this work, we investigate several photodetector designs based on $MoSe_2$-WSe_2 vdWHs through combined density functional theory (DFT) and quantum transport calculations. Typical device structures are shown in Fig. 1. Geometries with a partial and total TMDC overlap are considered, with a p-doped (n-doped) left (right) extension and an intrinsic region in the middle. We show that the partial overlap structure can enable a non-zero photocurrent even without built-in potential, in contrast to the full overlap one.

Method: All DFT calculations are performed with VASP using GGA-PBE [4]. The plane-wave Hamiltonian from VASP is then transformed into a basis of maximally localized Wannier functions (MLWFs) employing the wannier90 tool [5]. The resulting MLWF Hamiltonian is upscaled to a larger orthorhombic cell and repeated to construct the simulation domains with partial and total overlap.

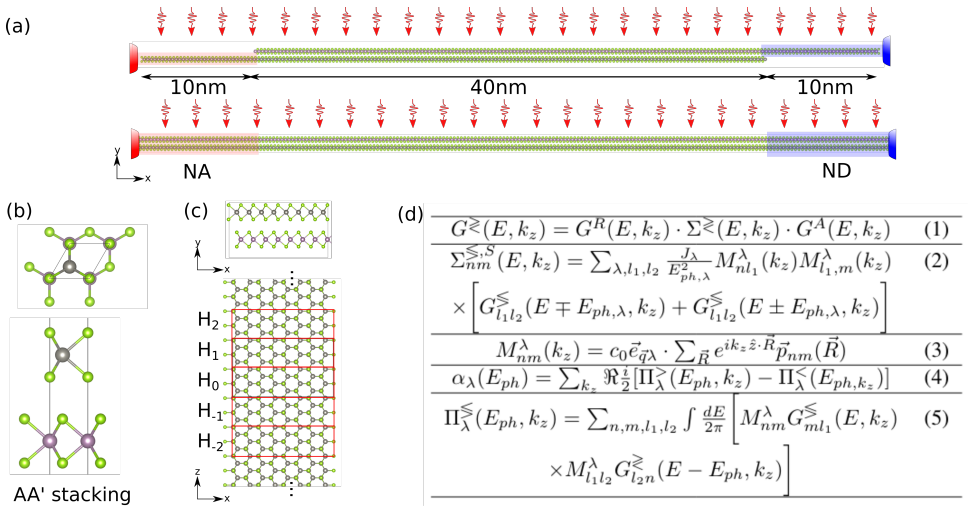
The quantum transport equations based on the Non-equilibrium Green's Function (NEGF) formalism are at the core of our device simulations [6]. They are summarized in Fig. 1. To take into account the electron-photon (e-phot) interactions, a dedicated scattering self-energy is implemented. The e-phot coupling matrix elements are obtained from the momentum operator $\vec{p}$ under the dipole approximation, which is computed in real space as the commutator between the Hamiltonian operator and the position operator. Both quantities are evaluated in the same MLWF basis. In our simulations, light enters the 2-D vdWHs orthogonal to the surface with a power density of J_λ. After solving the NEGF equations, the photo-excited current flowing through the vdWH device is computed from the $G^<$ with an atomistic resolution. These results can be used to evaluate the optical absorption coefficient α under the random phase approximation (see Fig. 1d).

Results: Fig. 2(a) shows the band structure of the $MoSe_2$-WSe_2 vdWH, Fig. 2(b) the absorption coefficient α as a function of the photon energy E_{ph}. It is reported for the $MoSe_2$ monolayer, WSe_2 monolayer, and the vdWH. We notice that the onset of α for the vdWH is lower than for both monolayers, indicating the presence of interlayer absorption.

Fig. 3 presents the quantum efficiency (QE) of different photodetector designs as a function of E_{ph} and the built-in potential V_{bi}. The QE is defined as the number of e-h pairs generating the electrical current I divided by the number of incident photons: *i.e.*, QE$= \frac{I/e_0}{LJ/E_{ph}}$, with $L = 60$ nm the total illuminated length. For the full overlap structure, the highest QE ($\approx$19%) happens at $E_{ph} = 1.9$ eV, and no photocurrent is generated with $V_{bi} = 0$. This is because the full overlap structure is symmetric under flat band condition: the current flows cancel each other, as shown by the spectral current density in Fig. 3(c). For the partial overlap structure, we find non-zero QE for $V_{bi} = 0$. To elucidate this phenomenon, we decompose the photocurrent flowing along the device into 4 intra- and inter-layer components shown in Fig. 4(a). By further separating the electron and hole currents in Figs. 4(b) and (c), we find out that the electron current is mostly located inside the $MoSe_2$ layer, while the hole current is equally distributed inside both layers. Fig. 4(d) presents the energy-resolved current density along the device illustrating the e-h pair generation.

Conclusion: In this work, we employ a recently developed electron-photon scattering model within the framework of NEGF to study different photodetector designs based on 2-D TMDC vdWHs. We demonstrate that the partial overlap structure can enable a good photo-responsivity even at zero build-in potential, thanks to the type-II band alignment and the generation of interlayer electron-hole pairs. This photodetector structure largely facilitates the fabrication since no gate nor chemical doping of the 2-D TMDC is required.

References: [1] C.-H. Lee, G.-H. Lee, A. M. Van Der Zande, *et al.*, Nature Nanotech. vol. 9, pp. 676–681, 2014. [2] Y. Jiang, S. Chen, W. Zheng, *et al.*, Light Sci. Appl. vol. 10, 72, 2021. [3] P. Rivera, K. L. Seyler, H. Yu, *et al.*, Science vol. 351, 688, 2016. [4] G. Kresse and J. Furthmüller, Phys. Rev. B vol. 54, 11169, 1996. [5] A. A. Mostofi, J. R. Yates, G. Pizzi, *et al.*, Comput. Phys. Commun., vol. 185, pp. 2309–2310, 2014. [6] A. Szabó, R. Rhyner, and M. Luisier, Phys. Rev. B, vol. 92, 035435, 2015.



$$G^{\gtrless}(E,k_z) = G^R(E,k_z)\cdot\Sigma^{\gtrless}(E,k_z)\cdot G^A(E,k_z) \quad (1)$$

$$\Sigma^{\gtrless,S}_{nm}(E,k_z) = \sum_{\lambda,l_1,l_2}\frac{J_\lambda}{E^2_{ph,\lambda}} M^\lambda_{nl_1}(k_z) M^\lambda_{l_1,m}(k_z) \quad (2)$$

$$\times\left[G^{\gtrless}_{l_1l_2}(E\mp E_{ph,\lambda},k_z)+G^{\gtrless}_{l_1l_2}(E\pm E_{ph,\lambda},k_z)\right]$$

$$M^\lambda_{nm}(k_z) = c_0\vec{e}_{\vec{q}\lambda}\cdot\sum_{\vec{R}} e^{ik_z\hat{z}\cdot\vec{R}}\vec{p}_{nm}(\vec{R}) \quad (3)$$

$$\alpha_\lambda(E_{ph}) = \sum_{k_z}\Re\tfrac{i}{2}[\Pi^>_\lambda(E_{ph},k_z)-\Pi^<_\lambda(E_{ph,k_z})] \quad (4)$$

$$\Pi^{\gtrless}_\lambda(E_{ph},k_z) = \sum_{n,m,l_1,l_2}\int\tfrac{dE}{2\pi}\left[M^\lambda_{nm}G^{\gtrless}_{ml_1}(E,k_z)\right. \quad (5)$$

$$\left.\times M^\lambda_{l_1l_2}G^{\gtrless}_{l_2n}(E-E_{ph},k_z)\right]$$

Fig. 1: (a) $MoSe_2$-WSe_2 vdW photodiodes with partial (top) and full (bottom) overlap. The left (right) 10-nm region is doped with an acceptor (donor) concentration of N_A (N_D). (b) Unit cell of the AA'-stacked $MoSe_2$-WSe_2 vdWH. (c) Orthorhombic unit cell used in transport simulations. (d) Summary of quantum transport equations with electron-photon scattering.

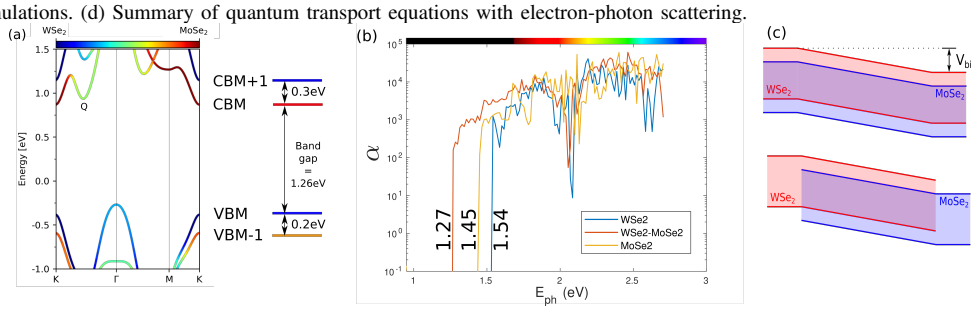


Fig. 2: (a) Partial bandstructure of an AA'-stacked $MoSe_2$-WSe_2 vdWH projected onto the orbitals of $MoSe_2$ (red) and WSe_2 (blue). The top bar indicates the orbital projection weight. (b) Calculated absorption spectra of the $MoSe_2$-WSe_2 vdWH and of its constituent monolayers. (c) Sketch of the conduction and valence band energies along the device. V_{bi} denotes the built-in potential.

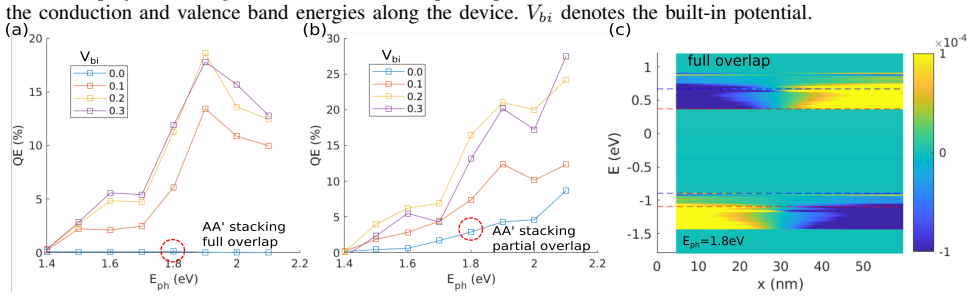


Fig. 3: (a) Quantum efficiency (QE) of a $MoSe_2$-WSe_2 vdWH photodiode with a full overlap geometry for photon energies $E_{ph} = 1.4-2.2$ eV and built-in potential $V_{bi} = 0-0.3$ V. (b) Same as (a), but for the partial overlap structure. (c) Spectral photo-current density at E_{ph}=1.8 eV and V_{bi}=0 V for the structure with total overlap. The dashed lines represent the CBM+1, CBM, VBM, and VBM-1 bands (see Fig. 2(a)).

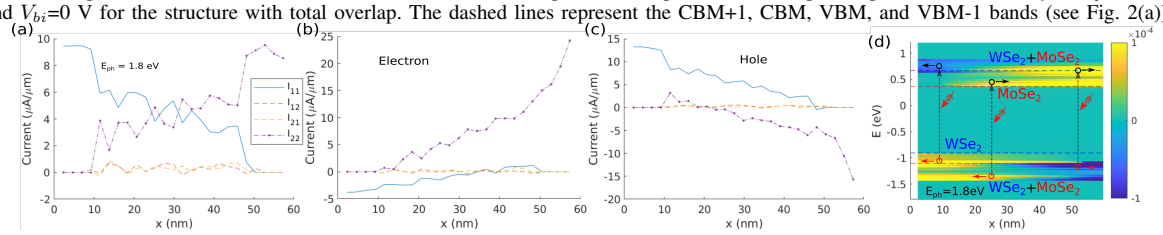


Fig. 4: (a) Photo-current flowing through the $MoSe_2$-WSe_2 photodiode with partial overlap at $V_{bi} = 0$, and E_{ph}=1.8 eV. I_{11} (I_{22}) is the current flowing between atoms in the WSe_2 ($MoSe_2$) layer, while I_{12} (I_{21}) is the inter-layer current flowing from the WSe_2 ($MoSe_2$) layer to the $MoSe_2$ (WSe_2) layer. (b) Same as (a), but only for electrons. (c) Same as (a), but only for holes. (d) Spectral photo-current density. The dashed lines represent the CBM+1, CBM, VBM, and VBM-1 bands.

Characterization and Modeling of Drain Lag using a Modified RC Network in the ASM-HEMT Framework

Mohammad Sajid Nazir[1], Ahtisham Pampori[1], Raghvendra Dangi[1], Pragya Kushwaha[2], Ekta Yadav[2], Santanu Sinha[2], and Yogesh Singh Chauhan[1]

[1]*Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India, 208016.*

[2]*Micro Electronics Group, Space Applications Centre, Indian Space Research Organisation, Ahmedabad, India*

Introduction

Gallium Nitride (GaN) based High Electron Mobility Transistors (HEMTs) show promising features and have been a subject of interest for several decades. Although significant efforts have been made to improve the reliability [1], these devices still exhibit current degradation due to charge trapping. In this paper, we characterize a GaN HEMT device under multiple pulsed conditions and present an RC network based model, implemented in the industry-standard ASM-HEMT model [2], to capture trapping effects and their self-limiting behavior.

Device Characterization

To observe current degradation with respect to time, quiescent voltages are set at (0V, 0V) and the device is measured at a bias of (-2.5V, 10V) over a pulse width of 20μsec at a 20% duty-cycle. To study the impact of drain induced trapping, pulsed characterization is performed at a fixed gate quiescent of -7V with varying drain quiescent bias. For these measurements, a pulse width of 500ns is applied at a 0.5% duty-cycle.

Results and Discussion

With an increasing drain quiescent voltage, the device exhibits a higher dynamic ON-resistance (R_{ON}) and a significant positive shift in threshold voltage (V_{OFF}). The change in threshold voltage can be attributed to the increased electric field under the gate due to a high drain bias, coupled with the trapping of electrons near the gate-drain edge due to high fields [3]. While the use of RC networks is a common approach to model trapping related current degradation [4, 5, 6], a model that takes the self-limiting behaviour of traps into account is still missing in the literature. The self-limiting behavior implies a condition where a trapped carrier generates an electrostatic potential that acts as a barrier for the incoming electrons, resulting in a decreased trapping probability [7]. Equivalent RC networks used to account for trapping without taking the self-limiting behaviour into account may not represent the variation of threshold voltage properly and can easily overestimate the dynamic R_{ON}. In our model, we introduce this effect using a negative feedback $R_{LE} = 1 - \alpha_{FB} V_{TRAP}$. Also, following from [8], drain bias induced space charge formation in the substrate can act as a parasitic back-gate (V_{SC})– degrading the current by reducing the 2-DEG [9], and simultaneously changing the rate of trapping. This effect is included in the model using a variable resistor R_{DE}=$R_D * NS0ACCD/(NS0ACCD - n_{SC})$, where $NS0ACCD$ is the drain access region concentration and $n_{SC} = V_{SC} * C_B/q$. The RC network model already present in ASM HEMT [4, 10] is modified to include these effects by using R_{LE} and R_{DE} as shown in Fig. 1(a). Fig. 1(b) shows the impact of R_{LE} on V_{TRAP} generated over time – as V_{TRAP} increases, negative feedback reduces R_{LE} thus increasing current flow through R_{LE} and subsequently limiting the increase of V_{TRAP}. R_E is kept at a higher value than R_T (emission time > capturing time) and a diode is used to ensure different paths for the charging and discharging of C_T. Validation of the proposed network for multiple quiescent conditions is shown in Figs. 1(c), 2, 3 and 4. Voltage dependent current sources are used (Eq. (1) and Eq. (2)) and effective trap potentials used to modulate ASM-HEMT parameters given by Eq. (3) and Eq. (4).

$$I_{SC} = \alpha_{SC} * V(d) \tag{1}$$

$$I_{TRAP} = \alpha_1 \sqrt{\alpha_2 V(d)^{\alpha_3}} \tag{2}$$

$$Voff_1 = RC1 * V_{TRAP} \tag{3}$$

$$Voff_2 = RC2 * (V_{TRAP} + V_{sc}) \tag{4}$$

Extraction Flow

With the quiescent condition of ($V_{GSQ} : -7V, V_{DSQ} : 0V$) chosen as reference, the initial set of ASM-HEMT parameters is extracted. The quiescent condition of ($V_{GSQ} : -7V, V_{DSQ} : 25V$) is then used to estimate the final values of trap parameters (α_{SC},α_1,α_2,α_3,α_{FB},RC1,RC2), keeping the basic parameters obtained in the first extraction unchanged. To simplify the extraction procedure, the trap potential is stored in two variables $Voff_1$ and $Voff_1$. $Voff_1$ is used to modulate the parameter VOFF and $Voff_2$ to emulate the effect of trapping on DIBL, mobility, saturation velocity and access region parameters.

Conclusion

We presented the characterization and empirical modeling of drain lag effects in AlGaN/GaN HEMTs. The proposed model works well under all quiescent conditions and can be extended for gate-lag by changing the voltage dependence of the current sources from drain to gate.

References

[1] D. Bisi et al., in IEEE TED, 60, 10, 3166-3175, 2013
[2] S. Khandelwal et al., in IEEE TED, 66, 1, 80-86, 2019
[3] I. P. Singh et al., in IEEE TED, 68, 2, 503-509, 2021.
[4] S. Agnihotri et al., IEEE INDICON, 2015, 1-4, 2015.
[5] J. Couvidat et al., IEEE/MTT-S IMS, 720-723, 2018.
[6] Jardel et al., MTT, IEEE Tran. 55. 2660 - 2669, 2007.
[7] T. Wosi´nski, J. Appl. Phys., 65, 4, 1566–1570, 1989.
[8] Sghaier et al., MJ, 37, 4, 363-370, 2006.
[9] Wang et al., JECS. 13, 872-876, 2014.
[10] S. Aamir Ahsan et al., in IEEE JEDS, 5, 5, 310-319, 2017.

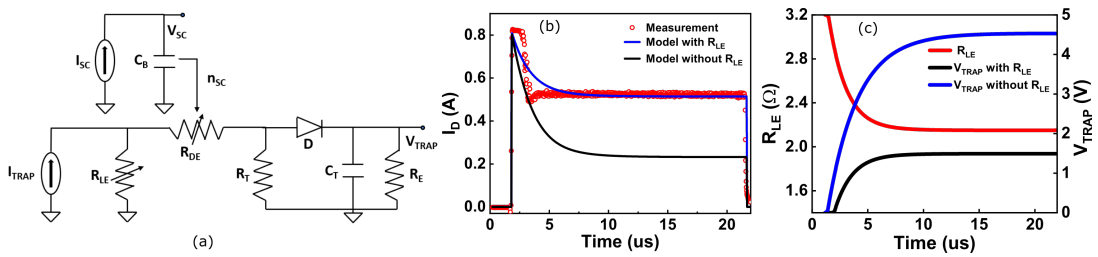


Fig. 1: (a) RC network used to implement trapping behaviour (b) Effect of trapping on drain current over a pulse width of $20\mu s$ (c) Impact on trap potential with and without feedback network

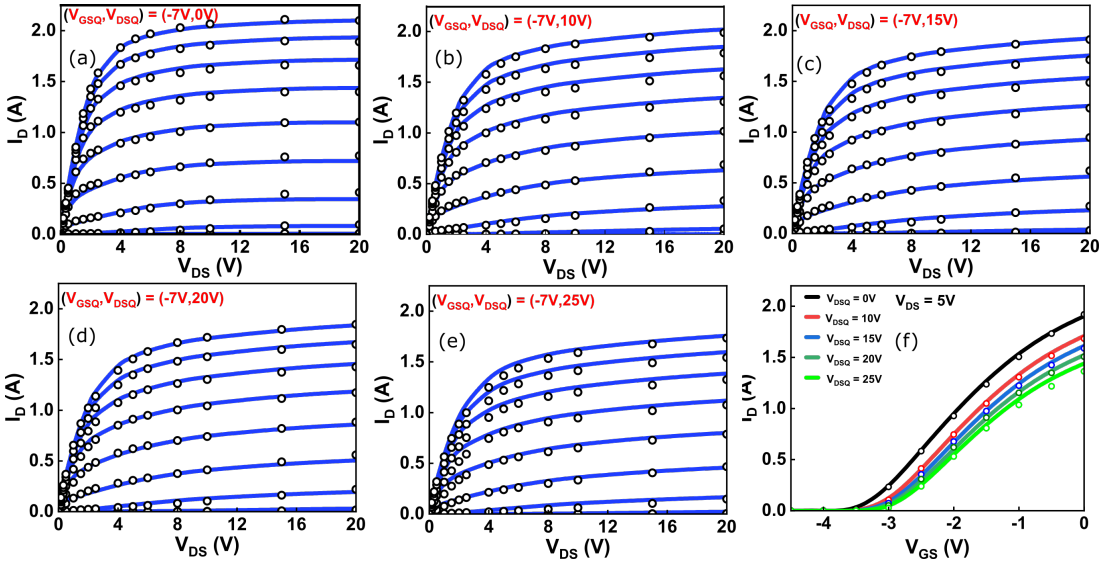


Fig. 2: Measurement (Symbols) and Simulated (Solid lines) (a)-(e) Pulsed output characteristics for the following quiescent bias points: $(V_{GSQ}, V_{DSQ}) = (-7V, 0V), (V_{GSQ}, V_{DSQ}) = (-7V, 10V), (V_{GSQ}, V_{DSQ}) = (-7V, 15V), (V_{GSQ}, V_{DSQ}) = (-7V, 20V), (V_{GSQ}, V_{DSQ}) = (-7V, 25V)$ (f) Drain current vs. Gate voltage at V_{DS}=5V for quiescent biases shown in legends.

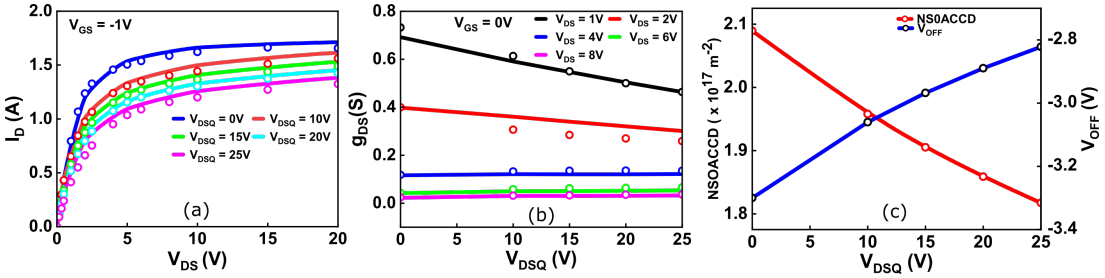


Fig. 3: (a) - (b) Measurement (symbols) and simulated (solid lines) (a) Drain current vs. Drain Voltage for quiescent conditions shown in legends (b) Dynamic conductance variation with drain quiescent condition (c) Access region 2-DEG and threshold voltage variation with drain quiescent condition.

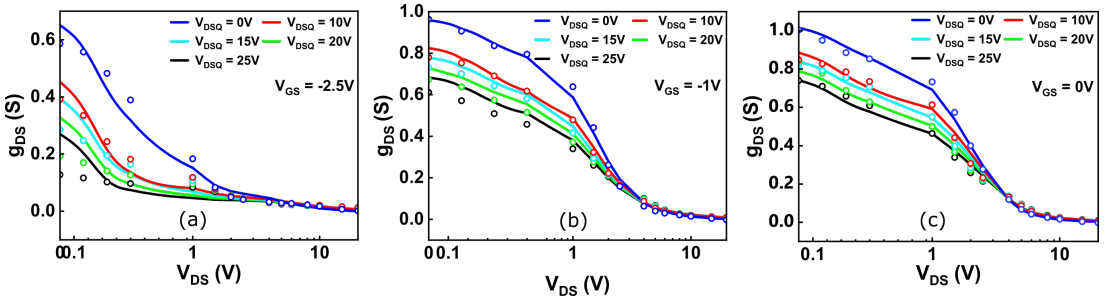


Fig. 4: Measurement (Symbols) and Simulated (Solid lines) (a) - (c) Dynamic conductance of device at gate voltage (-2.5V, -1V and 0V) for quiescent conditions shown in legends.

Efficient and accurate defect level modelling in monolayer MoS_2 via GW+DFT with open boundary conditions

Guido Gandus[1,2], Youseung Lee[1], Leonard Deuschle[1], Daniele Passerone[2], and Mathieu Luisier[1]
[1]Integrated Systems Laboratory, ETH Zürich, Switzerland; [2]nanotech@surfaces, EMPA, Switzerland

Introduction: The physical dimension of Si logic transistors are approaching the atomic limit, thus requiring novel architectures and/or high-mobility channel materials for future technology nodes. Logic switches based on two-dimensional (2D) transition-metal dichalcogenide (TMD) monolayers have thus been proposed to continue Moore's scaling law, thanks to their remarkable electronic properties. However, several works [1,2] reported that various defects inside these monolayers may limit their performance as logic devices, mainly through charged impurity scattering and defect-induced trap levels. In particular, the "mid-gap" states introduced by those impurities are presumably at the origin of large Schottky barriers (SB) and high contact resistances. Therefore, in order to understand the physics related to defects in 2D TMD monolayers and to guide device design, *ab initio* simulations are required. In this work, we propose an efficient GW algorithm combined with density functional theory (DFT) to accurately describe defect levels in 2D TMD monolayers. In conventional GW calculations, environmental effects from substrates are included to obtain the realistic bandgap of 2-D monolayers while requiring huge computational resources [3]. Our method, so-called projected p-GW, overcomes this issue by projections onto a defect subspace while removing spurious interactions between periodic images by means of open boundary conditions, as illustrated in Fig. 1. This algorithm can correctly predict the position of defect levels in the bandgap while ensuring the efficiency by resorting to the DFT-level bandgap. We then apply this method to the most common defect in MoS_2 monolayers: S vacancy.

Algorithm: Our algorithm is summarized in Fig. 2. We consider a central region containing a defect and consisting of integer repetitions of a unit cell called "principal layer" (PL) [4,5]. Our GW self-energy $\mathbf{\Sigma_{GW_\Delta}}$ is only calculated for a region surrounding the defect, and then coupled to pristine MoS_2 along all semi-infinite directions. The procedure articulates itself around three steps: 1) a DFT calculation of the central region builds the Hamiltonian of the defect+MoS_2 system at a mean-field level; 2) a boundary self-energy $\mathbf{\Sigma_B}$ replaces the periodic boundary conditions (PBCs) of DFT; 3) projection onto an orthogonal subspace surrounding the defect defines the GW region.

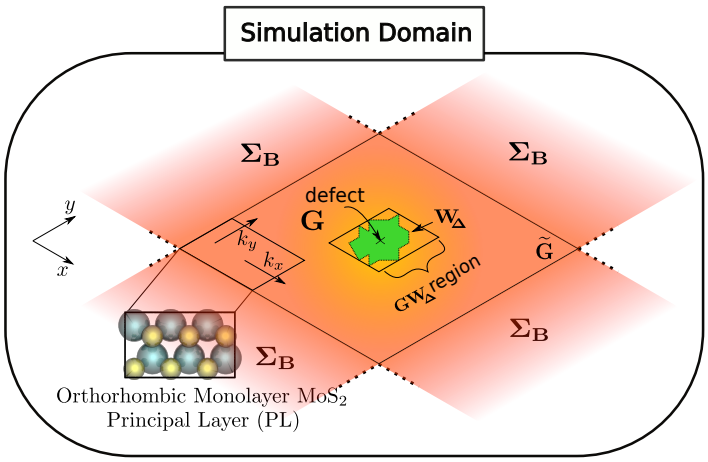
We consider a central region which is large enough to safely assume that the potential at the boundaries is converged to the one of bulk MoS_2. $\mathbf{\Sigma_B}$ can then be efficiently obtained from a k-point calculation of the PL by extending our procedure in Ref. [4], as summarized in the orange box of Fig. 2. This efficient and precise approach allows us to treat the system as "open" and effectively simulate the defect as isolated. Indeed, this avoids unwanted interferences or bound state patterns related to the PBCs.

The projection is performed onto a subspace of the central region $\mathbf{\Delta}$ that contains the defect and is orthogonal to the rest of the system [6]. The screened interaction $\mathbf{W_\Delta}$ calculated from this subspace is then multiplied by the Green's function of the defect to obtain $\mathbf{\Sigma_{GW_\Delta}}$.

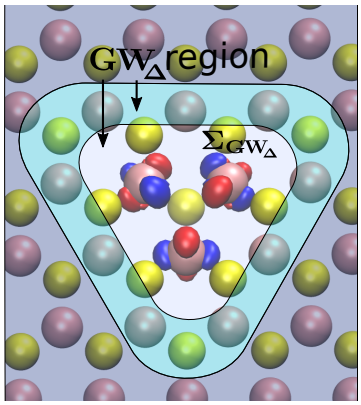
Results: We study the effect of S vacancies in 2D MoS_2 monolayers. The central region is composed of 4×6 repetitions of a PL composed of 6 Mo and 12 S atoms as shown in Fig. 1a. The electronic structure calculation of the PL is over-sampled with a $11 \times 12 \times 1$ k-mesh to obtain a $\mathbf{\Sigma_B}$ that precisely describes the bulk MoS_2 states. The GW region is shown in Fig. 1b together with the wavefunction of the state created by the defect. For $\mathbf{W_\Delta}$ we take into account up to the 2nd nearest neighbor to the vacancy, i.e. 12 Mo and 13 S atoms. The defect state is essentially a superposition of the 3d Mo orbitals closest to the vacancy, which allows us to define the defect as the 3 Mo and their surrounding S atoms. $\mathbf{\Sigma_{GW_\Delta}}$ in then computed for this region only. We calculate the density-of-states (DOS), the projected DOS (PDOS) and the electron transmission and report these results in Fig. 3. It is apparent from the DOS and the PDOS that the effect of the many-body correction is to shift the energy levels of the defect while preserving the DFT properties, i.e. the bandgap, as also corroborated by the conservation of the bulk-like electronic transmission. Previous k-point GW studies of full defect+MoS_2 structures found similar positions of the defect level with respect to the corresponding band edges [3]. This indicates that our p-GW algorithm can accurately predict trap-levels with minimal computational burden.

Conclusions: We proposed a novel algorithm to locally and efficiently apply many-body corrections using GW to a region surrounding a defect. Periodic self-interactions are removed by virtue of an efficient boundary self-energy calculation. The presented algorithm is then applied to S vacancy defects in a MoS_2 monolayer. Our method is a first step toward an inclusion on many-body methods beyond DFT in large scale simulations of realistic devices.

References: [1] Rai, A. et al., *Materials Today*, 339–401 (2022) [2] Lee, et al., *IEDM*, 24.4 (2019). [3] Naik, Mit H. et al., *Phys. Rev. Mater.*, 2, 084002 (2018). [4] Gandus, G. et al., *SISPAD*, 177 - 180 (2020). [5] Papior, N. et al., *Phys. Rev. B* 100, 195417 (2019). [6] Jacob, D. et al., *J. Condens. Matter Phys.*, 27(24) (2015).



(a) Schematic view of a defected region.



(b) GW region for a S vacancy.

Fig. 1: (a) Schematic view of a defected region connected to semi-infinite leads. The interaction with the leads is described by Σ_B. Electron-electron interactions are treated in a subspace denoted GW region. (b) GW region considered in this work to model S vacancies in MoS_2 monolayers. The isosurfaces represent the wavefunction of the state created by the defect.

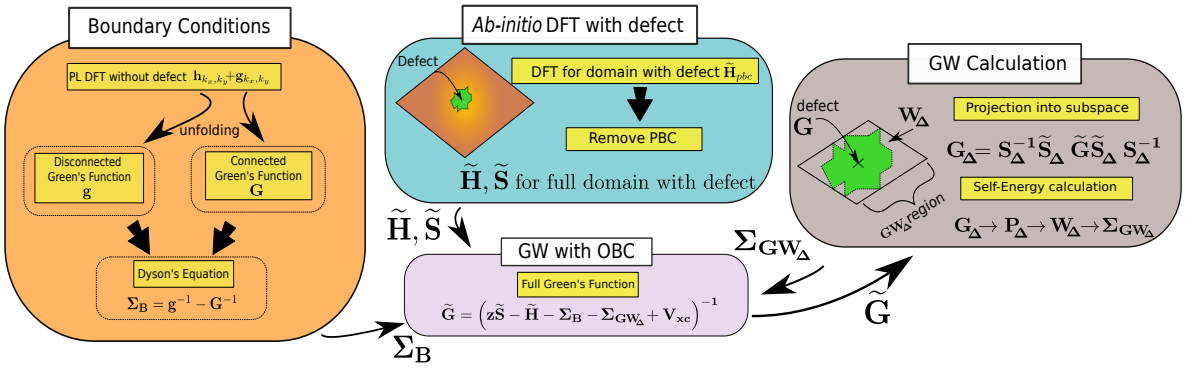
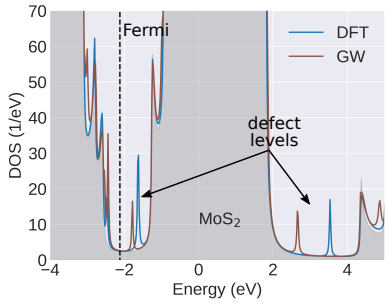
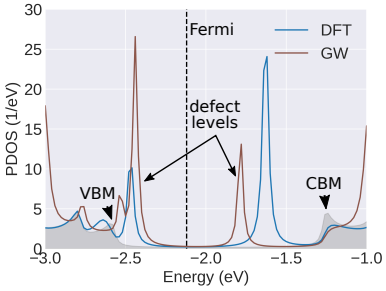


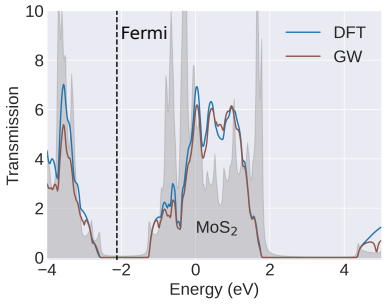
Fig. 2: Flowchart of the p-GW method for efficient modelling of defected structures with GW corrections. (Orange box) Σ_B is constructed from a DFT calculation of a periodic PL. (Blue box) The $\tilde{H}$ and $\tilde{S}$ of the defected region are obtained from a separate DFT calculation by removing the PBCs. (Gray box) Σ_{GW_Δ} is computed for a subspace containing the defect where G is constrained to the defect and W_Δ includes its surroundings. (Purple box) Equation for the full Green's function coupling all boxes.



(a) Density-of-states.



(b) Projected density-of-states.



(c) Transmission function.

Fig. 3: Results for S vacancy obtained by DFT and the proposed p-GW method. The total density-of-states of the central region (a) and projected onto the GW subspace (b) show that the many-body correction shifts the defect level states while maintaining the DFT bulk properties. (c) Transmission function through the defected structure. The onset of the electron transmission around the fundamental gap is also preserved by the p-GW correction.

Prediction of the evolution of defects induced by the heated implantation process: Contribution of kinetic Monte Carlo in a multi-scale modeling framework

P.L. Julliard*[1,2], A. Johnsson[3], R. Demoulin[2], R. Monflier[2], A. Jay[2],
D. Rideau[1], P. Pichler[3], A. Hémeryck[2], and F. Cristiano[2]
[1]STMicroelectronics, Crolles, France
[2]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France
[3]Fraunhofer Institute for Integrated Systems and Device Technology (IISB), Erlangen, Germany
Email: pierre-louis.julliard@st.com

***Abstract*—The extended defects formed as a result of heated implantation and thermal annealing are studied using transmission electron microscopy and Kinetic Monte Carlo simulations. We highlight the relevance of using Kinetic Monte Carlo to provide information for continuum scale simulations and also the value of integrating molecular dynamics results for its calibration.**

***Index Terms*—Kinetic Monte Carlo, Process simulations, Heated implantation, Multi-scale modeling.**

I. Introduction

Leakage current in transistors is known to increase in presence of extended defects such as {311} and dislocation loops [1]. Extended defects can be related to the implantation and annealing process. Thus, innovative process conditions to reduce defects concentration, such as heated implantation, are currently explored [2]. To assist in the optimization of the heated implantation process, Kinetic Monte Carlo (KMC) is an efficient methodology as it takes temperature into account and is integrated in state-of-the-art Technology Assisted Computer Design (TCAD) software [3]. In this work, the KMC as implemented in [3] is used to investigate the defects remaining after the heated implantation and annealing process, from small interstitial clusters to the formation of dislocation loops (DLs). The simulation results are compared with transmission electron microscopy (TEM) for three implantation conditions where the temperature varies. The experimental results (see section II) reveal two distinct defect formation regimes for which it is necessary to use different levels of modeling as a function of temperature (see section III), *i.e.* a hybrid approach coupling the process KMC and continuum solvers of [3] for the low implantation temperatures and, for high implantation temperature, a full KMC scheme.

II. Experimental results

We describe the results obtained by heated implantation for Si wafers implanted with Arsenic. Three wafers were implanted with Arsenic at RT, 150 °C and 500 °C. For each of these temperatures, the implantation is performed in two steps: a first implantation at 180 keV with a dose of 10^{14} cm^{-2} and a second implantation at 100 keV with a dose of 8×10^{13} cm^{-2}. An identical annealing treatment was performed for the three implanted wafers (see Tab. I). These as-implanted wafers are then analyzed by TEM and after the annealing cycle, as shown in Fig. 1.

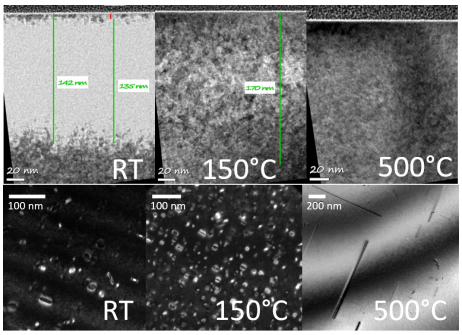


Figure 1: TEM analysis of As implanted Si wafers at three temperatures: RT, 150 °C and 500 °C (Top) cross-section images carried out in as-implanted Si wafers. (Bottom) plane-view images after annealing as given in Tab. I.

In Fig. 1-(Top), describing the material after implantation, we observe amorphization layer at RT, and a damaged material at 150 °C and 500 °C. After annealing (Fig. 1-(Bottom)), for both RT and 150 °C implantations, DLs are observed in TEM. Implantation at high temperature (500 °C) exhibits a different kind of defects, *i.e.* extended {311} defects.

Table I: Annealing sequence following the implantations at RT, 150 °C and 500 °C.

Annealing steps	1	2	3	4	5	6
Temperature (°C)	625	750	700	625	800	750
Time (minutes)	120	60	210	52	30	60

III. Simulations results

The continuum simulation model implemented in [3] is consistent with TEM results for RT implantation. In the case of 150 °C and 500 °C implantations using the software [3] the KMC model reproduces the TEM observation of an absence of amorphous layer (not shown) whereas the continuum model implemented

Table II: Interstitial densities trapped in DLs.

I in loops (cm^{-2})	TEM	full KMC	Continuum	Hybrid
RT	1.4×10^{14}	1.2×10^{14}	1.3×10^{14}	1.8×10^{14}
150 °C	1.6×10^{14}	2.3×10^{14}	1.6×10^{14}	3.0×10^{14}

(without additional parameter calibrations) does not reproduce it. In this paper, we focus on the simulation of defects at the end of the annealing sequence. To replicate the two observed experimental trends at the end of annealing (DLs in RT and 150 °C implantations and {311} in 500 °C implantation), we use two modeling approaches: a coupling between the process KMC and continuum solvers of [3] is proposed for RT and 150 °C implantation temperatures (see section III-A). For high temperature implantation (500 °C), a full KMC simulation is performed using molecular dynamics (MD) simulation results [4] for a detailed calibration of interstitial clusters (see section III-B).

A. Using hybrid KMC for RT and 150 °C implantations

The KMC simulates DLs after the annealing in good consistency with the TEM images. KMC provides an estimation of density of interstitials trapped in DLs close to the TEM count for RT implantation but it overestimates it by 50% for the 150 °C implantation (Tab. II). Using TEM images and additional calibrations to estimate excess interstitials after the implantation, a continuum simulation using the model of [5] provides a simulation consistent with TEM interstitial count (Tab. II).

To take advantage of the speed of continuum simulations without having to calibrate its parameters from TEM experiments, a hybrid approach combining KMC for implantation and continuum simulations for annealing was performed. The excess interstitials present in the amorphous pockets (intersitials-vacancies clusters) and interstitial clusters in the KMC simulations are converted into a continuous data field associated to I_2. The KMC is used for the implantation and for the beginning of the annealing sequence, until there is no more amorphous pockets in the simulation domain. Here, the interstitial densities in DLs simulated are of the same order of magnitude than in the TEM count, the densities of interstitials in DLs in the 150 °C being overestimated. The use of the KMC-continuum switch divides the computational cost time by a factor of 8 compared to a full KMC simulation using the same area.

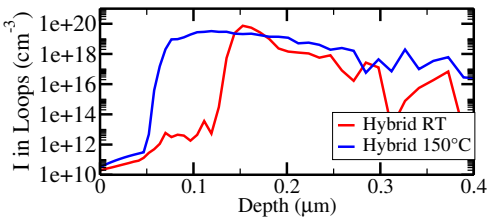


Figure 2: Number of I trapped in DLs using the hybrid approach combining KMC and continuum model after implantation and amorphous pockets disappearance at RT (red) and 150 °C (blue).

B. Dependence of SMIC size for 500 °C implantations

In the 500 °C implantation case, the KMC predicts the formation of DLs, result which is in disagreement with the {311} defects observed in TEM images. In the 500°C implant case, the annealing sequences being the same for the three wafers, the difference which leads to {311} formation must originate from the implantation step. During the 500 °C implantation the interstitials form stable interstitial clusters (SMICs) while for RT and 150 °C, the interstitials are in much larger amorphous pockets or clusters. The SMICs have been the subject of several atomic investigations in the literature [4][6][7]. Two kinds of SMICs were identified: the chain-like defects type, precursor of {311} [7] and the Arai type whose structure is close to the one found in [6]. The second type of SMICs is known to have very stable structures for I_{4n}. An increase of the difference between I_{4n} and other SMICs in the activation energies for the emission of an interstitial from a given cluster is expected to slow down the growth of extended defects. Indeed, this modification enables to simulate {311} in 500 °C case (Fig. 3). Further investigations on the SMICs obtained after implant at 500 °C using MD simulations will be shown in the full-paper.

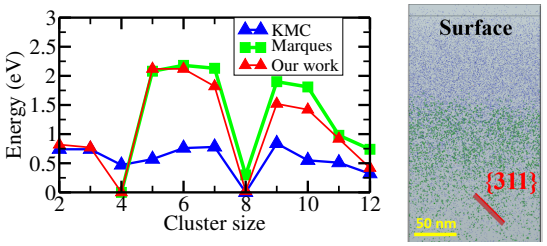


Figure 3: (Left) Difference in formation energies of clusters I_n and I_{n-1} as implemented in the KMC (blue), calculated for the Arai SMIC family using MD [4] (green) and as they were implemented in this work to simulate {311} defects in 500 °C implantation using KMC (red). (Right) {311} defect simulated in KMC using the energies in the left figures. Small green particles correspond to I_4 clusters and blue one to As atoms.

IV. Conclusion

The KMC demonstrates its validity in simulating heated implantation and reproducing experimental observations. For RT and 150 °C, a smart coupling of KMC with continuum model for the simulation of the annealing sequence allows to greatly improve its efficiency. For 500 °C, we show the importance of a fine calibration, based on atomic MD simulations to explain the observed formation of {311} defects.

References

[1] Nyamhere et al., "A comprehensive study of the impact of dislocation loops on leakage currents in si shallow junction devices," Journal of Applied Physics, vol. 118, no. 18, p. 184501, 2015.
[2] Wen et al., "Finfet io device performance gain with heated implantation," in 2018 22nd International Conference on Ion Implantation Technology (IIT). IEEE, 2018, pp. 106–109.
[3] Sentaurus Process User Guide, 2019.03, Synopsis Inc.
[4] L. A. Marques et al., "{001} loops in silicon unraveled," Acta Materialia, vol. 166, pp. 192–201, 2019.
[5] Wolf et al., "Modeling the annealing of dislocation loops in implanted c-si solar cells," IEEE Journal of Photovoltaics, vol. 4, no. 3, pp. 851–858, 2014.
[6] N. Arai, et al., "Self-interstitial clustering in crystalline silicon," Physical review letters, vol. 78, no. 22, p. 4265, 1997.
[7] J. Kim et al., "Stability of si-interstitial defects: From point to extended defects," Physical Review Letters, vol. 84, no. 3, p. 503, 2000.

Surface scattering impact on Si/$TiSi_2$ contact resistance

Kantawong Vuttivorakulchai*, Mohammad Ali Pourghaderi, Yoon-Suk Kim, Uihui Kwon, and Dae Sin Kim
Computational Science and Engineering Team, Innovation Center,
Device Solution Business, Samsung Electronics Co., Hwasung-si, Gyeonggi-do, Republic of Korea.
*Email Address: k.vuttivorak@samsung.com.

***Abstract*— This study reports the practical limit on contact resistance of Si (100)/$TiSi_2$ system. The maximum transmission in silicide contact is estimated via overlap of conducting modes. This ideal limit is then compared versus the coherent transport, where DFT-NEGF is applied to a pool of model interfaces. It is shown that interface scattering increases the contact resistance by an order of magnitude. The barrier tunneling strongly depends on surface termination and coordination defects. For the studied samples, the trend of contact resistance matches with transmission at Fermi level.**

***Keywords—Contact resistance, transmission, surface scattering, valley filtering, DFT-NEGF*.**

I. Introduction

The scaling of transistors in advanced nodes has effectively decreased the channel resistance. This makes the parasitic contributions particularly important, as they do not scale easily. Perhaps, the most challenging component is the contact resistance. Currently, titanium silicide ($TiSi_2$) is the common metal choice. In this study, some basic aspects of carrier transport in Si/$TiSi_2$ system are explored.

The samples of Si/C49-$TiSi_2$ junctions are prepared in (100) orientation. For this orientation, $TiSi_2$ may have 5 different terminations: three Si-rich and two Ti-rich atomic planes. Figure 1 shows the corresponding hetero-interfaces for all $TiSi_2$ terminations. The density functional theory (DFT) relaxation is performed under PseudoDojo pseudopotentials [1] with PBEsol functional and double-zeta polarized basis. The density cut-off energy is set to 600 Ha. The ***k***-grids of 3×3×1 are applied for the density calculations. The contact resistance is calculated in two different ways: valley filtering and DFT non-equilibrium Green's function (DFT-NEGF) method. For the valley filtering, our procedure is similar to the work in Refs. [2] and [3]. After structural relaxation, the unit cells of bulk-representative parts are used for band structure calculation with uniform ***k***-grids of 101×101×101. The ***k***-resolved distribution of modes, $M(E, \boldsymbol{k})$, is then extracted for the relevant energy range. From the valley filtered $M(E, \boldsymbol{k})$, the contact resistance (ρ) is computed according to Landauer formalism [3].

For DFT-NEGF method, meta-GGA exchange-correlation functional of Tran & Blaha (TB09) is used [4]. The c-parameter is set to 1.03511. The density mesh cut-off is set to 100 Ha. Following the experimental reports [5], silicon is doped to 3×10^{20} cm^{-3} level with donors. The transmission spectrums are computed with uniform ***k***-grids of $51\times51\times1$ and used for contact resistance calculations [6].

II. Results and Discussions

Figure 2 shows ***k***-resolved conducting modes of C49-$TiSi_2$, Si, and the corresponding overlap. The conducting modes are effectively filtered in high energy range as it is shown in Fig. 3(a). Consequently, the difference of intrinsic and valley-filtered resistance is increased at high carrier concentration, shown in Figs. 3(b) and 3(c). Though these results consider the non-ideality of metal and mismatch of symmetries, the reflection and tunneling resistance at interface is neglected. Therefore, these results can set the lowest theoretical limit for contact resistance. To investigate the impact of interface scattering, DFT-NEGF simulations are carried out. The corresponding ***k***-resolved transmissions are shown in Fig. 4. The transmissions across hetero-interface are much lower than ideal limit in Fig. 2. In particular, the conducting path near Γ-point is totally vanished for all cases. This is mainly due to the heavy effective mass in longitudinal direction. Figure 5(a) demonstrates the comparison of $T(E)$ among five different interfaces. As shown in Fig. 5(b), the contact resistances are much larger than theoretical limit in Fig. 3(c), i.e. $5.9\times10^{-11}\ \Omega\cdot cm^2$. Interestingly, the atomic structure in case 3 gives the highest resistance. This is due to the fact that the specific Si-termination does not produce the metallic character of $TiSi_2$. As a result, the tunneling path between semiconductor- and first metallic-plane gets longer, which effectively hamper the tunneling and transmission. In general, the combination of bond-intimacy and coordination defects will determine the level of tunneling. The effective contribution of these factors can be presented by transmission at Fermi level, $T(E_f)$. As shown in Fig. 5(c), there is strong correlation between contact resistance and $1/T(E_f)$. Consequently, $T(E_f)$ can be used as a metric for fast screening of model interfaces.

III. Conclusion

The impact of interface chemistry on tunneling current in silicide contact is investigated. Among the prepared samples with similar Schottky barrier heights, those with shorter tunneling path and coordination defects closer to Fermi level produce lower contact resistance. For the nominal doping level, the tunneling current is much lower than theoretical limit. This suggests that there is a big room to engineer the interface and optimize the contact resistance.

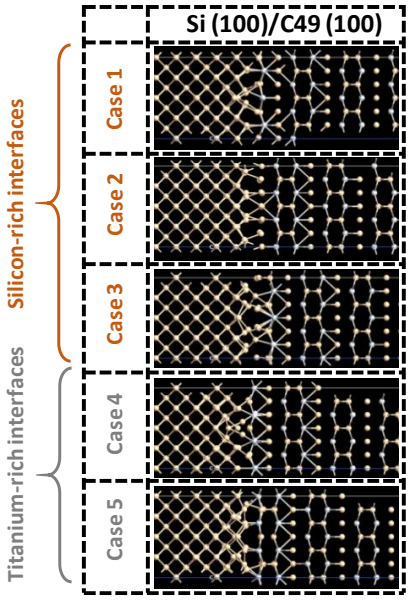


Fig. 1. The atomic structures of Si (100)/$TiSi_2$ (100) interfaces.

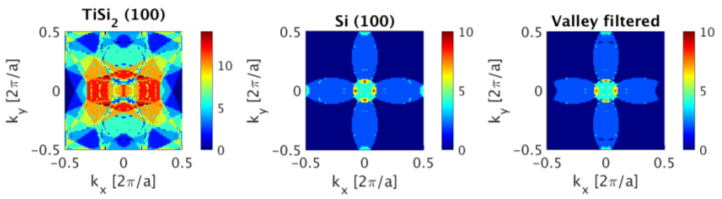


Fig. 2. The transverse momentum-resolved distribution of modes (DOM) of $TiSi_2$, Si with surface orientation in (100) direction, and its valley filtered DOM, respectively. Here, the energy level is at 0.08 eV from conduction band minimum.

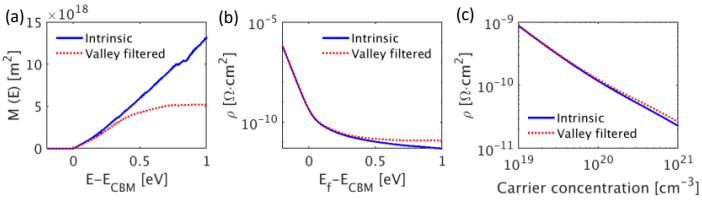


Fig. 3. The comparison between the intrinsic limit properties of Si and its valley filtered by $TiSi_2$ at 300 K. (a) The available conducting modes per unit area, $M(E)$ as function of energy from the conduction band minimum (CBM). (b) The resistance at different Fermi level from CBM. (c) The resistance as function of electron concentration.

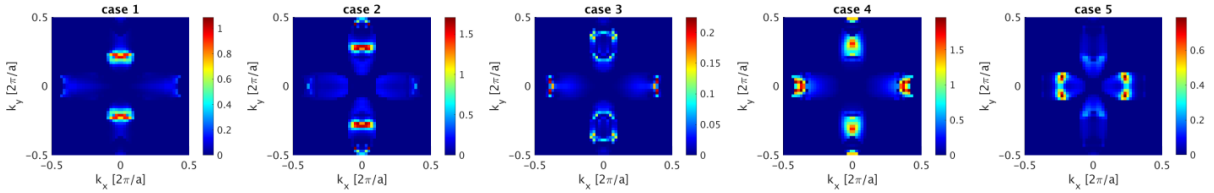


Fig. 4. The DFT-NEGF calculations of the transverse ***k***-resolved transmission spectrums of Si (100) /$TiSi_2$ (100) interfaces as shown in Fig 1.

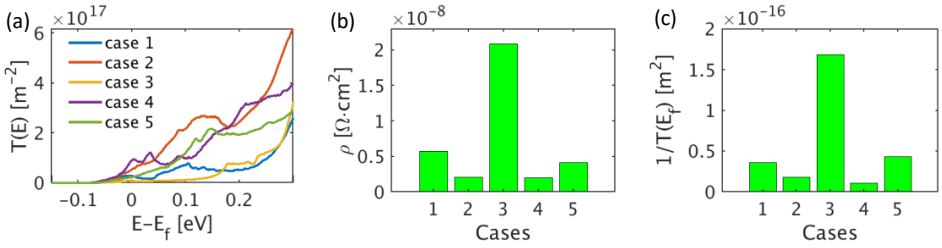


Fig. 5. (a) The DFT-NEGF transmissions per unit area of 5 cases of Si (100) doped by 3×10^{20} cm^{-3} /$TiSi_2$ (100) interfaces (Fig. 1) as function of energy from its Fermi level. (b) The corresponding resistances at room temperature. (c) The inverse transmission at Fermi level of these 5 cases.

References

[1] M. J. van Setten *et al.*, "The PseudoDojo: Training and Grading a 85 Element Optimized Norm-conserving Pseudopotential Table," *Comput. Phys. Commun.*, vol. 226, pp. 39–54, May 2018.

[2] G. Hegde and R. C. Bowen, "Effect of Realistic Metal Electronic Structure on the Lower Limit of Contact Resistivity of Epitaxial Metal-Semiconductor Contacts," *Appl. Phys. Lett.*, vol. 105, pp. 053511, August 2014.

[3] J. Maassen *et al.*, "Full Band Calculations of the Intrinsic Lower Limit of Contact Resistivity," *Appl. Phys. Lett.*, vol. 102, pp. 111605, March 2013.

[4] S. Smidstrup *et al.*, "QuantumATK: An Integrated Platform of Electronic and Atomic-scale Modelling Tools," *J. Phys.: Condens. Matter*, vol. 32, pp. 015901, October 2019.

[5] H. Yu *et al.*, "Titanium Silicide on Si:P with Precontact Amorphization Implantation Treatment: Contact Resistivity Approaching 1×10^{-9} Ohm-cm^2," *IEEE Transactions on Electron Devices*, vol. 63, pp. 4632–4641, December 2016, doi: 10.1109/TED.2016.2616587.

[6] T. Markussen and K. Stokbro, "Metal-InGaAs Contact Resistance Calculations from First Principles," *2016 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2016, pp. 373–376, doi: 10.1109/SISPAD.2016.7605224.