2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) | 979-8-3315-1635-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/SISPAD62626.2024.10733041

Diffusion-Based Machine Learning Method for Accelerating Quantum Transport Simulations in Nanowire Transistors

Preslav Aleksandrov
Department of Computer Science
University of Cambridge
Email: pa511@cam.ac.uk

Pranav Acharya
James Watt School of Engineering
University of Glasgow
Email: 2670931A@student.gla.ac.uk

Vihar Georgiev
James Watt School of Engineering
University of Glasgow
Email: vihar.georgiev@glasgow.ac.uk

***Abstract*—Numerical device simulations of nano-scale transistors are a vital part of semiconductor research. Recent advancements in machine learning are slowly finding their way into the nanoelectronic device simulation community. Machine learning (ML) supported methods hold the promise of significantly reducing the cost and wall-clock time of simulations.**

In this work, we present a new way to utilise advanced image processing machine learning techniques to accelerate quantum transport simulations of nanowire transistors. Our method uses a ML based diffusion model to accelerate the speed of the numerical simulations and reduce the self-consistent numerical iterations. The ML diffusion model is based on the popular UNet architecture together with application specific modifications, which we have developed for the purpose of our research. Our ML based method improves the speed and the performance of our in-house code, called NESS, by up to 60%.

I. Introduction

Nanowires play a pivotal role in the advancement of nanoelectronics, offering unique properties that are essential for the development of high-performance electronic devices [1]. These one-dimensional structures are crucial in applications ranging from transistors to sensors, where their quantum mechanical properties can be leveraged for enhanced functionality [2]. Hence, accurate and efficient simulation methods are critical for exploring and optimising nanowire-based devices in order to decrease fabrication time and cost [3].

One way to reduce the simulation time and increase the speed of the simulations is to combine numerical simulations with ML methods. For example, our previous work introduced the ML-NEGF method, which utilised convolutional neural networks (CNNs) to accelerate non-equilibrium Green's function (NEGF) simulations [4]. While this approach significantly improved convergence speed, there remained room for further enhancements. In this follow-up study, we present a novel model architecture that replaces CNNs with diffusion models, which have demonstrated superior performance in capturing complex physical phenomena [5]. Diffusion models, with their ability to better model stochastic processes, provide a more accurate and efficient framework for our simulations. Additionally, we have developed an improved location map, which significantly enhances the spatial resolution and accuracy of our simulations. This advancement allows us to significantly reduce simulation time by paving the way for more effective exploration and development of the next-generation nanoelectronic devices.

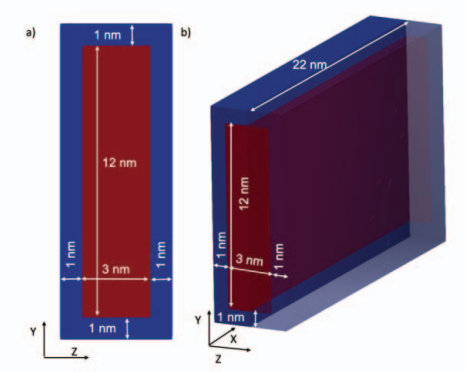


Fig. 1: Schematic representation of the nanowire structure. (a) Cross-sectional view showing the dimensions: 12 nm height, 3 nm width, and 1 nm gate oxide thickness on each side. (b) Three-dimensional view of the nanowire, illustrating its 22 nm length. The red region represents the silicon semiconductor channel, while the blue regions indicate the gate silicon oxide surrounding the channel.

II. Device Structure

To validate our new simulation methodology, we have designed a nanosheet transistor structure relating to the so-called 3 nm technology node [6]. Figure 1 illustrates the geometry created using our in-house structure generator implemented in NESS [7]. The device features a N-type gate-all-around (GAA) configuration with a Si channel length of 16 nm and source/drain lengths of 3 nm each, resulting in a total device length of 22 nm. The rectangular channel cross-section measures 3 nm x 12 nm, with a 1 nm thick SiO_2 layer. The channel is P-doped at $1e15cm^{-3}$, while the source and drain regions are N-doped at $1e20cm^{-3}$.

The process began with the design of various transistor geometries using the NESS structure generator, focusing on

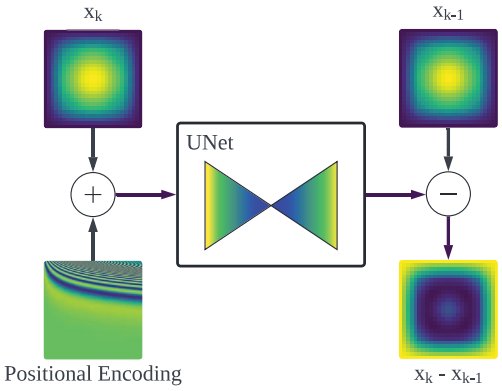


Fig. 2: Schematic view of a single step in the stable diffusion process. The UNet model takes as input the noisy image x_k at iteration k, along with a positional encoding. It predicts the noise residual $(x_k - x_{k-1})$, which is then used to estimate the less noisy image x_{k-1} for the previous timestep. This process is repeated iteratively to progressively denoise the image.

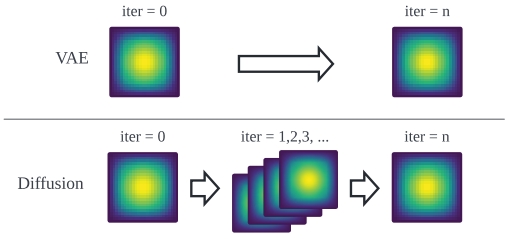


Fig. 3: Comparison of the Variational Autoencoder (VAE) and diffusion models. The top row illustrates the VAE process, which directly maps between the original image (iter = 0) and the latent representation (iter = n). The bottom row shows the diffusion process, which gradually adds noise to the image over multiple iterations (iter = 1, 2, 3, ..., n), creating a sequence of increasingly noisy versions of the original image. This step wise process allows diffusion models to learn a more detailed mapping between the image and noise spaces.

configurations relevant to the 3 nm node technologies and beyond [6]. To ensure the robustness and versatility of our model, the dataset includes variations in material properties, such as different oxide materials and their thicknesses, in order to evaluate their impact on device performance. Geometrical variations, including different channel lengths, widths, and doping profiles, are also incorporated. Simulations are conducted under various gate and drain biasing conditions to capture a comprehensive set of performance metrics, including current-voltage characteristics, 3D electron density and 3D potential profile distributions.

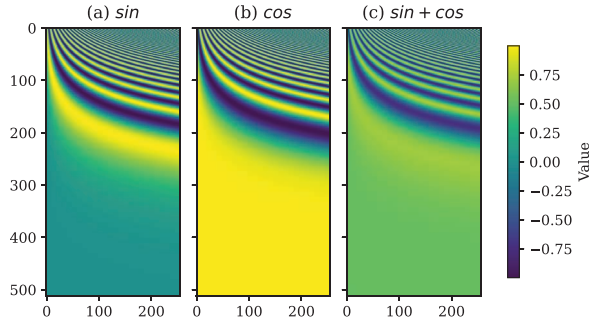


Fig. 4: Visualisation of 2D positional encoding for image-based applications. The figure shows three maps representing different components of the positional encoding scheme: (a) Sine encoding, (b) Cosine encoding, and (c) Combined sine and cosine encoding. The colour scale on the right indicates encoding values. This scheme creates a unique encoding for each pixel position, allowing the model to capture and utilise precise spatial information within images.

III. Machine Learning Model

In this work, we present a new method (dNEGF) that combines a NEGF (quantum mechanical) solver implemented in our code called NESS and a latent diffusion model shown in Figure 2. In contrast with other ML approaches available to expedite the quantum NEGF simulations, dNEGF executes iteration-based training. A comparison of dNEGF with conventional variational autoencoder (VAE) methods is shown in Figure 3. Iteration based training is one of the core components of the diffusion methods [5].The method operates by assuming that the difference between iteration in NEGF is predictable and follows known distribution. This idea is similar to simulating noise in conventional vision based ML. It has been shown that it is easier to predict the difference between two levels of noise, rather than remove the noise entirely [8]. Therefore, we have adopted the similar approach for our dNEGF methods. Using as a stepping stole our previous work, we employ a positional embedding [4]. However, ML literature suggests that linear positional maps underperform in more complex methods such as those shown in [9]. Figure 4 visualises the 2D sinusoidal positional encodings used in our dNEGF model. It is an adaptation of the similar model shown in [9].

Combining the advanced positional encoding techniques with the diffusion model aims to produce greater generality, transferability and expand the model applicability to an arbitrary simulation domain. Our model is used as a 'better' initial guess for the Poisson-NEGF self-consistency loop in NESS leading to a reduction of the simulation time. A vital part of the model is its fully convolutional architecture that allows transferability of the model to different device geometries.

The model, shown in Figure 2, uses the popular UNet [10] architecture used in other methods, such as stable diffusion [11]. The UNet architecture trains effectively on a few input images and outperforms sliding window convolutional

networks, such as the one used in [4]. The model input is a 7-channel image generated from information produced by an initial NEGF iteration. The first two channels are standardised slices of potential and of the logarithm of charge. The following two channels are the drain and gate voltages. The final three channels are the location maps.

IV. Simulation Methodology

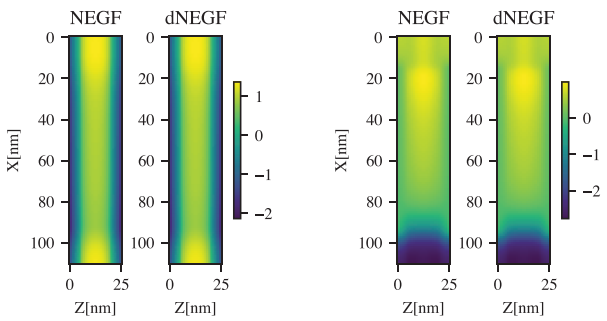
Diffusion models, which are particularly effective at capturing complex stochastic processes, provide a more accurate representation of the quantum mechanical behaviours in nanoelectronic devices [5] if compared to encoder methods. By leveraging the inherent strengths of diffusion models, our approach ensures faster convergence and greater precision in simulation outcomes. An integral component of our methodology is the refined location mapping technique. This technique significantly improves the spatial resolution of the simulations, enabling a more detailed and accurate depiction of the nanosheet transistor structure. The enhanced location map facilitates precise tracking of electron transport, charge and potential distribution within the device.

A. Simulation Process

1) **Structure Generation:** The simulation begins with the creation of the transistor geometry using the NESS structure generator [7]. The designed device features a gate-all-around configuration, with specific dimensions and doping concentrations tailored to reflect advanced 3 nm node technologies.
2) **Initial Conditions:** The initial conditions for the simulation are set, including doping levels, material properties, and boundary conditions. These parameters are crucial for accurately modelling the behaviour of the nanosheet transistor.
3) **Diffusion Model Application:** The diffusion models are then applied to simulate the quantum transport phenomena within the device [5]. These models excel at handling the non-linearities and stochastic nature of electron transport, providing a robust framework for the simulation.
4) **Iterative Convergence:** The simulation proceeds iteratively, with the diffusion models driving the convergence process. Each iteration refines the results, ensuring that the final output accurately represents the physical behaviour of the device.

V. Results

The results of our new dNEGF simulation methodology, which incorporates diffusion models and enhanced location mapping, demonstrate significant advancements in accuracy and numerical efficiency. For example, the diffusion models provide a more precise representation of quantum mechanical behaviours, resulting in improved current-voltage characteristics, electron density distributions, and potential profiles. Enhanced location mapping significantly boosts spatial resolution, allowing for a more detailed and accurate depiction of the transistor structure. These improvements collectively lead to a substantial reduction in computational time, enabling faster numerical convergence without sacrificing accuracy.



(a) Normalised charge (b) Normalised potential

Fig. 5: Comparison of normalised charge and potential distributions from the 'standard' NEGF and proposed dNEGF methods: (a) 2D normalised charge distribution; (b) 2D normalised potential distribution. In both sub-figures, results from the 'standard' NEGF method (left) and our newly developed dNEGF approach (right) are shown. The virtually identical distributions for both charge and potential validate the physical accuracy of our dNEGF method, demonstrating that it successfully reproduces the results of the 'standard' NEGF method.

The enhanced convergence of dNEGF, compared to the 'standard' NEGF methods, is demonstrated in Figure 6. From this figure it is clear that in all-gate biases the dNEGF model requires less interactions in order to reach the convergence criteria in comparison to the 'standard' NEGF method. On average, for all reported gate biases, the speed of the Poisson-NEGF self-consistent loop is increased by up to 60%.

Also, the physical validity of our methods is proven by comparing the 2D and 3D potential distribution between the dNEGF and standard NEGF methods. This is shown in Figure 5 as a cross section of the nanosheet transistor fields (potential and charge) and in Figure 7 as a comparison of the current-voltage characteristics. From Figure 5 it is clear that the 2D cross-section distribution of the charge and the potential is identical for both methods. This proves that the physical parameters, such as electrostatics and quantum mechanics, are identical. Hence, as it is expected, the current-voltage characteristics for both methods must be identical and indeed this is proven in Figure 7. Both methods produce identical current-voltage characteristics that overlap perfectly.

VI. Conclusion

In this study, we present an enhanced model architecture designed to accelerate the non-equilibrium Green's function (NEGF) simulations, building upon our previous ML-NEGF method [4]. This new approach (dNEGF) incorporates diffusion models, leading to notable improvements in both convergence speed and simulation accuracy. Our findings demonstrate a significant reduction in computational time, on average

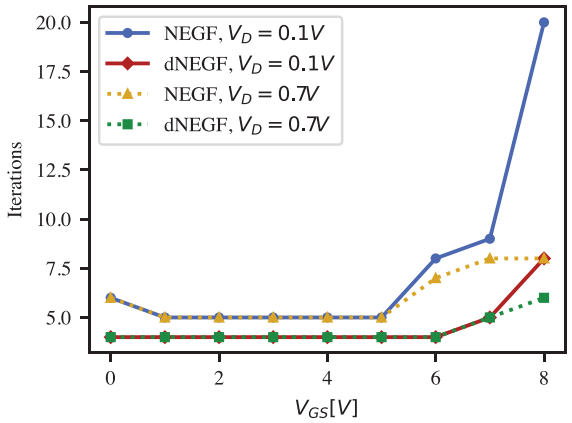


Fig. 6: Comparison of convergence between the 'standard' Non-Equilibrium Green's Function NEGF and dNEGF methods. The graph shows the number of iterations required to reach convergence for two different drain voltages ($V_D = 0.1V$ and $V_D = 0.7V$). The x-axis represents the gate-source voltage (V_{GS}) in volts, while the y-axis shows the number of iterations. For both low ($0.1V$) and high ($0.7V$) drain voltages, dNEGF consistently converges in fewer iterations if compared to the 'standard' NEGF method, demonstrating its improved computational efficiency.

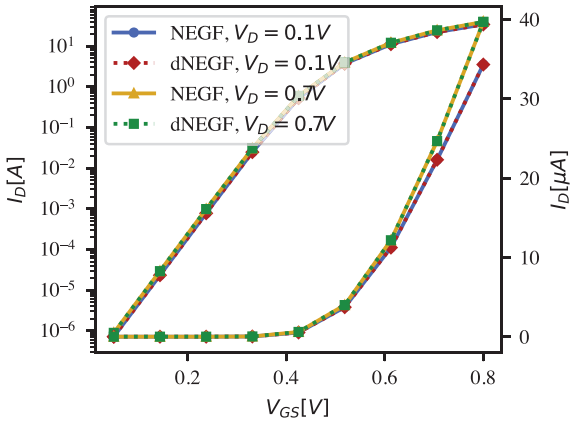


Fig. 7: Comparison of current-voltage (I_D-V_{GS}) characteristics obtained using the 'standard' Non-Equilibrium Green's Function (NEGF) and diffusion NEGF (dNEGF) methods. The graph shows drain current (I_D) versus gate-source voltage (V_{GS}) for two different drain voltages ($V_D = 0.1V$ and $V_D = 0.7V$). Both methods produce identical I_D-V_{GS} curves for both low and high drain voltages, demonstrating the accuracy of dNEGF if compared to the 'standard' NEGF method across different operating regimes.

60%, achieving faster modelling of quantum mechanical phenomena in silicon nanowire transistors.

The increased efficiency not only accelerates research timelines but it also reduces the computational resources required, making quantum mechanical simulations more energy efficient and less time consuming. Furthermore, the robustness of our enhanced model architecture suggests its potential applicability across a broader range of nanoelectronic devices and materials. This versatility is particularly important as the field of nanoelectronics continues to expand, encompassing new materials and innovative device architectures [1]. Future work will focus on further refining the model to enhance its performance and generality. We also plan to explore the integration of our approach with other simulation frameworks to extend its applicability.

REFERENCES

[1] Y. Cui, Q. Wei, H. Park, and C. M. Lieber, "Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species," *Science*, vol. 293, no. 5533, pp. 1289–1292, 2001. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1062711

[2] F. Patolsky, B. P. Timko, G. Zheng, and C. M. Lieber, "Nanowire-based nanoelectronic devices in the life sciences," *MRS Bulletin*, vol. 32, no. 2, p. 142–149, 2007.

[3] J. Wang, A. Rahman, A. Ghosh, G. Klimeck, and M. Lundstrom, "On the validity of the parabolic effective-mass approximation for the i-v calculation of silicon nanowire transistors," *IEEE Transactions on Electron Devices*, vol. 52, no. 7, pp. 1589–1595, 2005.

[4] P. Aleksandrov, A. Rezaei, T. Dutta, N. Xeni, A. Asenov, and V. Georgiev, "Convolutional machine learning method for accelerating nonequilibrium green's function simulations in nanosheet transistor," *IEEE Transactions on Electron Devices*, vol. 70, no. 10, pp. 5448–5453, 2023.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[6] K. J. Kuhn, "Considerations for ultimate cmos scaling," *IEEE Transactions on Electron Devices*, vol. 59, no. 7, pp. 1813–1828, 2012.

[7] S. Berrada, T. Dutta, H. Carrillo-Nunez, M. Duan, F. Adamu-Lema, J. Lee, V. Georgiev, C. Medina-Bailon, and A. Asenov, "Ness: new flexible nano-electronic simulation software," in *2018 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2018, pp. 22–25.

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 10 674–10 685. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01042