Title

A SPICE-Compatible Neural Network Compact Model for Efficient IC Simulations

ISBN

979-8-3315-1636-9

Authors

Tung, Chien-Ting

Salahuddin, Sayeef

Hu, Chenming

Publication Date

2024-01-27

DOI

10.1109/sispad62626.2024.10733248

A SPICE-compatible Neural Network Compact Model for Efficient IC Simulations

Chien-Ting Tung
Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA, USA
cttung@berkeley.edu

Sayeef Salahuddin
Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA, USA
sayeef@eecs.berkeley.edu

Chenming Hu
Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA, USA
hu@eecs.berkeley.edu

***Abstract*— We present a SPICE-compatible neural-network-based compact model (BSIM-NN) for advanced FETs. The model consists of an IV and QV network which have all terminal currents and charges and includes geometry dependence. The study of the activation functions is conducted to find the most efficient function for circuit simulations. The model is then implemented in Verilog-A with direct multiplication instead of loops to enhance the computational speed. We demonstrate and benchmark the neural network model performance in large circuit simulations with different network structures. It shows about 40 times speed improvement compared to the conventional compact model.**

Keywords—Compact model, machine learning, neural network, SPICE, Verilog-A

I. INTRODUCTION

Fast and accurate compact device models are crucial for IC design and technology development. Standard compact models, such as the Berkeley Short-channel IGFET Model (BSIM) [1, 2] uses physics-based equations or semi-empirical equations to model complex device phenomena. Developing accurate and computationally efficient formulas has become more and more challenging for advanced and emerging devices due to their complex physical effects such as quantum effect, and short channel effect [3].

Neural network/machine learning-based compact device models have gained much interest due to the potential to model complex device physics with high efficiency accurately. Several neural network-based compact models have been proposed and demonstrated the ability to accurately model the IV and CV characteristics, including geometry dependence or variability [4-11]. However, after implementation in Verilog-A, the NN models have often shown inferior simulation speed compared to conventional compact models [4-6]. This lack is typically attributed to the unavailability of matrix multiplication functions in Verilog-A and implementation in C has been suggested as a remedy. In this study, we show, by contrast, that a substantial speed increase over conventional compact models can be achieved with Verilog-A code itself, simply by avoiding loops and arrays. Indeed, a speed boost as much as ~40X is demonstrated.

II. MODEL FRAMEWORK

BSIM-NN consists of complete IV and QV networks which are trained with the measured/generated IV and CV characteristics with all terminal currents (I_D, I_G …) and charges (Q_G, Q_S, Q_D). In this work, the training and testing data are generated by a calibrated BSIM-CMG model card [4]. The geometry dependence (L, W, EOT …) is included in the training. The network structures and loss functions are shown in Fig. 1. For the IV network, the outputs are the transform of I_D and I_G by (1) and (2) so that the range of the data is easier to learn [4]. Second derivatives are included in loss functions (3) to improve accuracy [4, 5].

$$I_D = V_{DS} e^{y_1},\ y_1 = \ln\left(\frac{I_D}{V_{DS}}\right), \tag{1}$$

$$y_{2p} = \ln\left(\frac{I_G}{2} + \frac{\sqrt{I_G{}^2 + \Delta^2}}{2} + I_0\right),$$

$$y_{2n} = \ln\left(\frac{-I_G}{2} + \frac{\sqrt{I_G{}^2 + \Delta^2}}{2} + I_0\right), \tag{2}$$

$$loss = a \cdot RMS(y_1) + b \cdot RMS(g_m) + c \cdot RMS(g_{ds}) + d \cdot RMS(g_m') + e \cdot RMS(g_{ds}') + f \cdot RMS(y_{2p}) + f \cdot RMS(y_{2n}), \tag{3}$$

For the QV network, the outputs are the terminal charges (4).

$$y_{1,2,3} = Q_{G,S,D}, \tag{4}$$

The coefficients a-f are used to adjust the contribution of each term. We use an iterative training method to fine-tune the coefficients so that the magnitude of each term will be similar. During the training, we first set b-f to 0 and only trained with a. After that, we use the trained weights and biases to train the second time by setting b to be non-zero and so on. For each training, the magnitude of the new loss should be tuned to be comparable to the previous loss. Furthermore, for the QV loss function (5), an offset term ($Q_{G,S,D0}$) is introduced [10] to overcome the charge shift problem when training with only the capacitance data reported in [6]. $Q_{G,S,D0}$ are the charges at V_{GS}=V_{DS}=0V which can be set using physical estimation.

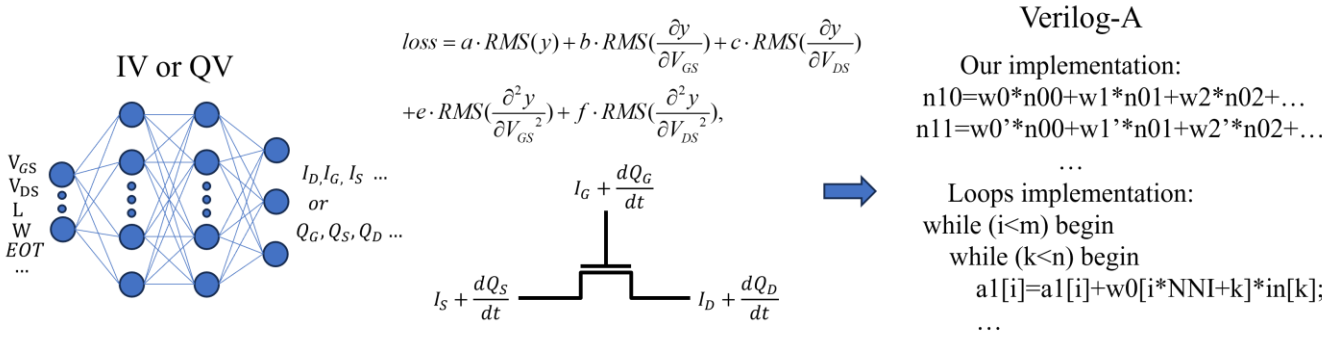

Fig. 1. The schematic diagram of the NN FET model, the loss function, and the Verilog-A implementation. The compact model has both IV and QV networks. The loss function includes up to second derivatives of outputs (y) and it can be adjusted by tuning the weights (a-f) of each term. Our implementation uses direct multiplication instead of loops.

$$loss = a \cdot RMS(\frac{\partial Q_{G,S,D}}{\partial V_{GS}}) + b \cdot RMS(\frac{\partial Q_{G,S,D}}{\partial V_{DS}}) + c \cdot RMS(\frac{\partial^2 Q_{G,S,D}}{\partial V_{GS}^2}) + d \cdot RMS(\frac{\partial^2 Q_{G,S,D}}{\partial V_{DS}^2}) + e \cdot RMS(Q_{G,S,D0}), \quad (5)$$

In this study, the number of hidden layers is fixed to 2. In addition to the network structure, the form of the activation function is also an important contributor to NN performance. Three functions are examined in this work: sigmoid, tanh, and ISRU ($x/\sqrt{1+x^2}$). Then, networks with these different activation functions are trained with 10 neurons in each hidden layer. The trained models are automatically coded into Verilog-A with a Python code that we developed. The weights, biases, and matrix multiplications are coded into direct multiplications element-by-element without using any array and loop in Verilog-A as shown in Fig. 1. This implementation style increases the SPICE simulation speed since the array and loop are inefficient in Verilog-A (Table I, Fig. 4-8). Fig. 2 & 3 show the fitting results of IV and CV characteristics with randomly selected bias, L, W, and EOT. We can achieve high accuracy with just 2 hidden layers and 10 neurons/hidden layers. Furthermore, the comparison of different activation functions using a 17-stage ring oscillator is summarized in Table I. All these NN models show a significant speed boost and similar iteration numbers compared to Verilog-A BSIM-CMG. Among the three activation functions, ISRU performs the best due to its simpler form and the lack of using an exponential function. A test result using loops for matrix multiplication is also shown and is much slower than direct multiplications. The number of multiplications and nonlinear functions in a two hidden-layer and 10 neurons/hidden-layer network should not be larger than the BSIM-CMG. However, the loop implementation result is slower than BSIM-CMG, confirming that direct multiplication is a more efficient approach in Verilog-A. In the following tests, ISRU is the activation function, and more comparisons between the implementations and network structures are presented.

	17-stage RO 50ns	Total iterations	Avg time per iteration
BSIM-CMG	116.53s	202802	5.745e-4s
sigmoid	19.99s	195486	1.023e-4s
tanh	16.97s	188670	8.996e-5s
ISRU	15.09s	193447	7.801e-5s
ISRU (loop)	172.41s	193327	8.918e-4s

Table. I. Comparison of BSIM-CMG and NN models with different activation functions using a 17-stage ring oscillator for 50 ns.

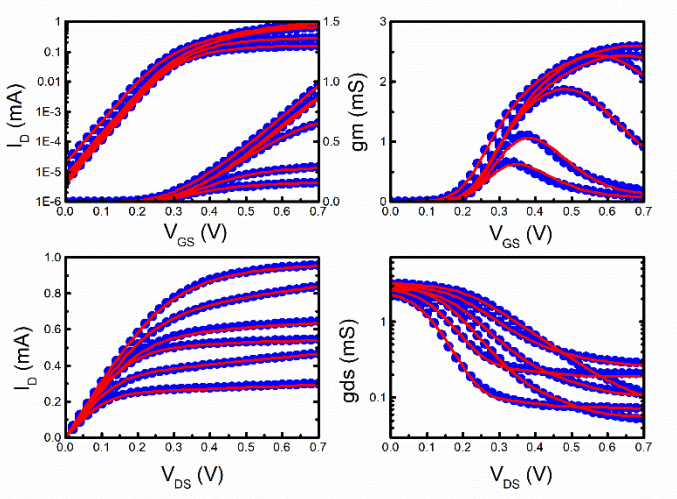

Fig. 2. The fitting results of IV characteristics of BSIM-NN. The symbols are the testing data and the lines are the model.

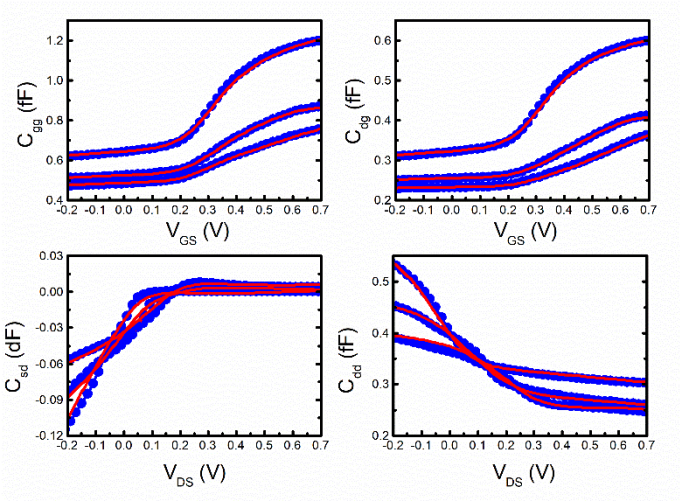

Fig. 3. The fitting results of CV characteristics of BSIM-NN. The symbols are the testing data and the lines are the model.

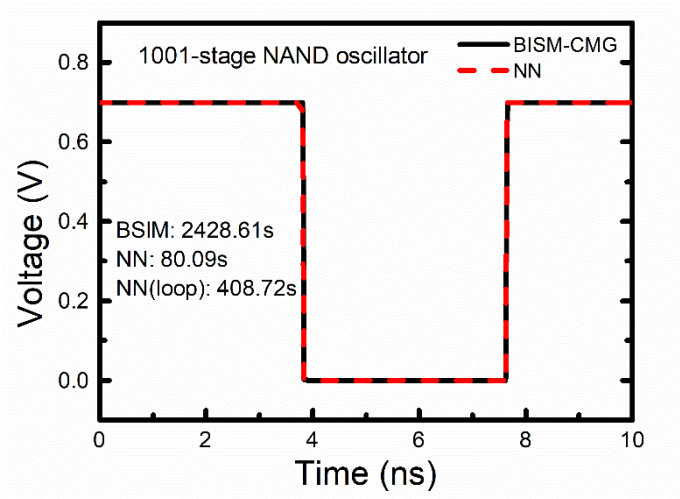

Fig. 4. Simulation of a 1001-stage NAND oscillator using BSIM-CMG and BSIM-NN.

III. Circuit Simulation and Benchmark

To demonstrate the model's robustness, several different circuits are tested. Fig. 4 shows the simulation result for a 1001-stage NAND oscillator compared to Verilog-A BSIM-CMG. The model is accurate with no convergence issues in larger circuits. Other circuits such as 16-bit full adder (Fig. 5) are also tested where results match BSIM-CMG well. In the above circuit, the BSIM-NN shows up to 30 times speed boost with our implementation. The loop implementations are slower in all of these cases.

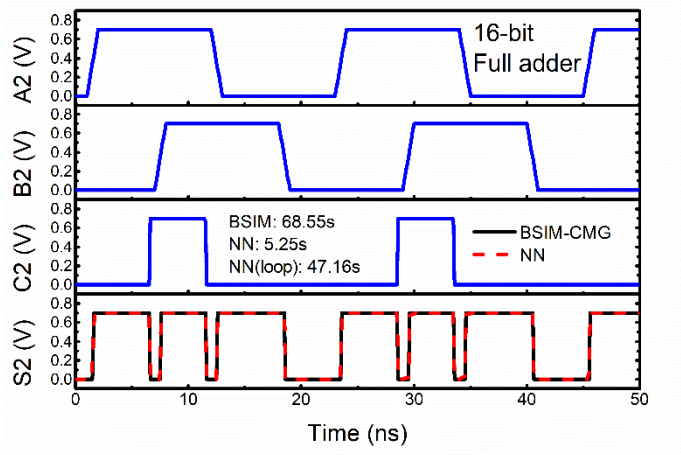

Fig. 5. Simulation of the 2nd bit in a 16-bit full adder using BSIM-CMG and BSIM-NN.

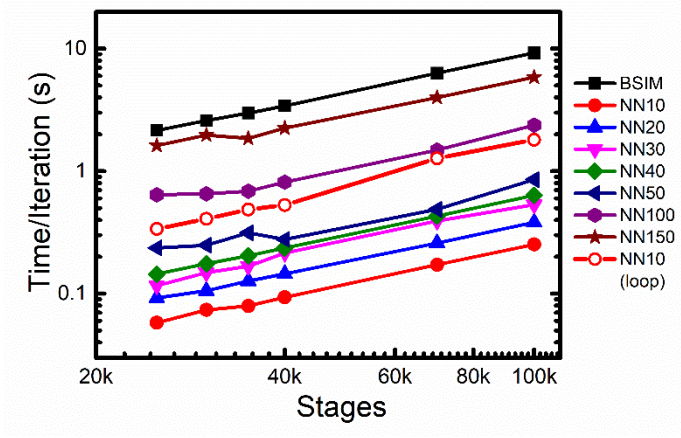

Fig. 6. The total simulation time of NAND oscillators from 25K+1 to 100K+1 stages. We test BSIM-CMG and NN models with different neuron numbers in the hidden layers.

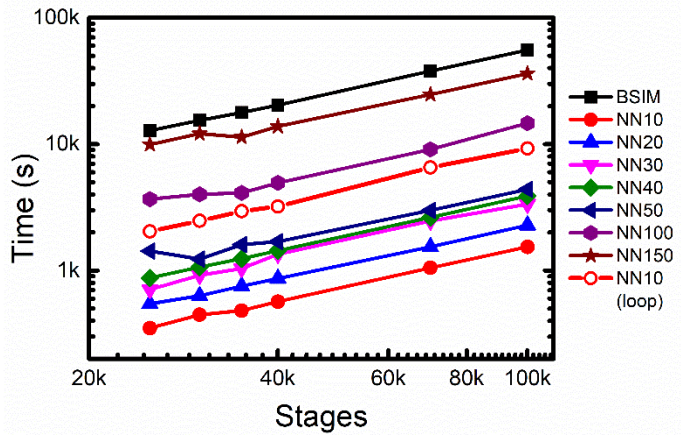

Fig. 7. The simulation time per iteration of NAND oscillators from 25K+1 to 100K+1 stages. We test BSIM-CMG and NN models with different neuron numbers in the hidden layers.

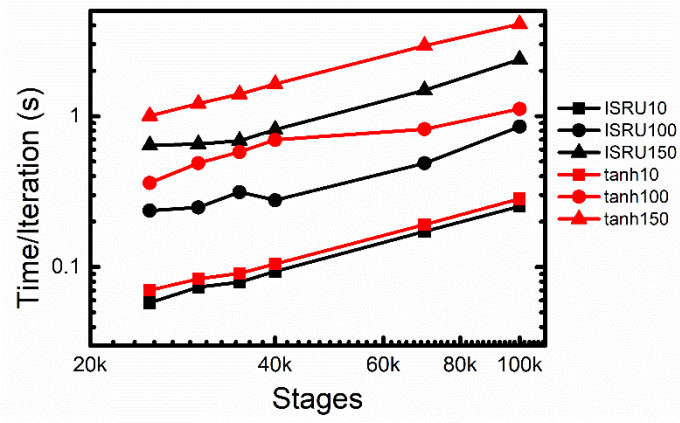

Fig. 8. The speed comparison of NN models with ISRU and tanh activation functions using NAND oscillators. 10, 100, and 150 are the neurons/hidden layers.

To examine the impact of the structure of NN models, we trained 7 different NN models with different numbers of neurons/hidden layers. They have 10, 20, 30, 40, 50, 100, and 150 neurons/hidden-layer in each network. An additional network with 10 neurons/hidden layers implemented with loop implementation is also tested (hollow symbol). These models and BSIM-CMG are used to simulate NAND oscillators from 25K+1 stages to 100K+1 stages. The testing result of total simulation time and time per iteration is shown in Fig. 6 & 7. All networks are faster than BSIM-CMG, and the smallest network is about 40 times faster in these large circuit simulations with comparable accuracy. Although in some cases, loop implementation can be faster than BSIM-CMG, it is still much slower than direct multiplications. Therefore, the Verilog-A NN models should be implemented with the direct multiplication approach. The test results suggest the potential of scaling NN models when more inputs come in. Importantly, for larger circuits, even the loop implementation is faster than BSIM-CMG, indicating that NN models' speed advantage increases in the more important case which is larger circuits (Fig. 4-7). Our tests also show that there is still some speed improvement for the networks with neurons greater than 100. This indicates the possibility of using a larger number of neurons to include more device parameters, biases, and temperature in the inputs and improve accuracy. Unlike standard models, all parameters are hard-coded, the NN models have the flexibility to choose a faster speed or more input parameters.

Finally, a test between models using ISRU and tanh with 10, 100, and 150 neurons/hidden layers is shown in Fig. 8. For all circuits, large and small, ISRU performs better than tanh.

IV. Conclusion

The number of transistors in integrated circuits is increasing rapidly, starting to exceed well beyond 100 billion. The time of simulation for such gigantic systems is poised to become a significant bottleneck. This work shows a pathway towards increasing simulation speed by as much as 40X by using neural networks while maintaining the ease and familiarity of Verilog-A implementation.

Acknowledgment

The authors gratefully acknowledge the funding support from Cadence, GlobalFoundries, Intel, Qualcomm, Samsung, ShanghaiTech, Synopsys, and TSMC through the Berkeley Device Modeling Center.

References

[1] J. P. Duarte *et al.*, "BSIM-CMG: Standard FinFET compact model for advanced circuit design," in *ESSCIRC Conference 2015 - 41st European Solid-State Circuits Conference (ESSCIRC)*, 14-18 Sept. 2015 2015, pp. 196-201, doi: 10.1109/ESSCIRC.2015.7313862.

[2] S. Khandelwal *et al.*, "BSIM-IMG: A Compact Model for Ultrathin-Body SOI MOSFETs With Back-Gate Control," *IEEE Transactions on Electron Devices*, vol. 59, no. 8, pp. 2019-2026, 2012, doi: 10.1109/TED.2012.2198065.

[3] G. Pahwa *et al.*, "Compact Modeling of Emerging IC Devices for Technology-Design Co-development," in *2022 International Electron Devices Meeting (IEDM)*, 3-7 Dec. 2022 2022, pp. 8.1.1-8.1.4, doi: 10.1109/IEDM45625.2022.10019433.

[4] C. T. Tung and C. Hu, "Neural Network-Based BSIM Transistor Model Framework: Currents, Charges, Variability, and Circuit Simulation," *IEEE Transactions on Electron Devices*, vol. 70, no. 4, pp. 2157-2160, 2023, doi: 10.1109/TED.2023.3244901.

[5] C. T. Tung, M. Y. Kao, and C. Hu, "Neural Network-Based I – V and C – V Modeling With High Accuracy and Potential Model Speed," *IEEE Transactions on Electron Devices*, pp. 1-4, 2022, doi: 10.1109/TED.2022.3208514.

[6] J. Wang, Y. H. Kim, J. Ryu, C. Jeong, W. Choi, and D. Kim, "Artificial Neural Network-Based Compact Modeling Methodology for Advanced Transistors," *IEEE Transactions on Electron Devices*, vol. 68, no. 3, pp. 1318-1325, 2021, doi: 10.1109/TED.2020.3048918.

[7] K. Sheelvardhan, S. Guglani, M. Ehteshamuddin, S. Roy, and A. Dasgupta, "Machine Learning Augmented Compact Modeling for Simultaneous Improvement in Computational Speed and Accuracy," *IEEE Transactions on Electron Devices*, pp. 1-7, 2023, doi: 10.1109/TED.2023.3251296.

[8] Y. S. Yang, Y. Li, and S. R. R. Kola, "A Physical-Based Artificial Neural Networks Compact Modeling Framework for Emerging FETs," *IEEE Transactions on Electron Devices*, vol. 71, no. 1, pp. 223-230, 2024, doi: 10.1109/TED.2023.3269410.

[9] Z. Yang, A. D. Gaidhane, K. Anderson, G. Workman, and Y. Cao, "Graph-Based Compact Model (GCM) for Efficient Transistor Parameter Extraction: A Machine Learning Approach on 12 nm FinFETs," *IEEE Transactions on Electron Devices*, vol. 71, no. 1, pp. 254-262, 2024, doi: 10.1109/TED.2023.3327973.

[10] C. T. Tung, S. Salahuddin, and C. Hu, "Non-Quasi-Static Modeling of Neural Network-based Transistor Compact Model for Fast Transient, AC, and RF Simulations," *IEEE Electron Device Letters*, vol. 45, no. 7, pp. 1277-1280, July 2024, doi: 10.1109/LED.2024.3404404.

[11] C. T. Tung, A. Pampori, C. K. Dabhi, S. Salahuddin, and C. Hu, "A Novel Neural Network-based Transistor Compact Model Including Self-Heating," *IEEE Electron Device Letters*, pp. 1-1, 2024, doi: 10.1109/LED.2024.3408151.