Physics-augmented Neural Compact Model for Emerging Device Technologies

Yohan Kim, Sanghoon Myung, Jisu Ryu, Changwook Jeong, and Dae Sin Kim
Data & Information Technology Center
Samsung Electronics
Hwaseong-si, Gyeonggi-do 18448, Korea
johnn.kim@samsung.com

***Abstract*— This paper proposes a novel compact modeling framework based on artificial neural networks and physics informed machine learning techniques. This physics-augmented neural compact model shows highly accurate fitting abilities and physically consistent inferences even at the unseen data. It is also scalable and technology independent, and consequently, is suitable for electrical modeling of new emerging devices. In addition, this neural compact model is able to cover both digital and analog circuit analysis due to the weight decay regularization as well as high order derivative losses. Finally, it is applied to promising DRAM and Logic technologies to be evaluated in terms of its scalability and fitting accuracy. The CMC's (Compact Model Coalition) standard model API (Application Programming Interface) supports the custom model implementation for SPICE. Therefore, this framework enables the circuit simulators to assess technology-independent PPA (Power, Performance, Area) and early-stage DTCO (Design Technology Co-optimization) for new emerging devices.**

Keywords—Compact Model, PDK, SPICE Circuit Simulation, Machine Learning, Artificial Neural Networks, Physical theory augmentation, Emerging device modeling, PPA and DTCO

I. INTRODUCTION

Over the past decade, CMOS technology is aggressively scaled down and the physical process limit is reached. As a result, various emerging devices with rapid changes in geometry, architecture, and material need to be evaluated through PPA assessments and DTCO activities in a timely manner. However, existing CMC's standard compact models are inappropriate for these emerging devices because developing new models is very time-consuming and strongly depends on each technology expertise. Therefore, we propose an artificial intelligence assisted compact model framework which is highly accurate, extremely general, and physics-embedded. It is going to change the paradigm of the expertise intensive model research and early-stage DTCO environments.

II. LIMITS OF THE EXISTING COMPACT MODELS

The existing compact models are at extreme ends of parametric (physics based) and non-parametric (empirical lookup table based) models. The analytical equations, represented by standard compact models, are physics-based and computationally efficient, however it takes very long time and requires special expertise to develop a reliable model for new technologies. In addition, the parametric analytical models support increasing number of empirical fitting parameters [1] even though they are built on solid physical theories because of various secondary effects and non-monotonic characteristics from modern complex process flows as shown in Fig. 1.

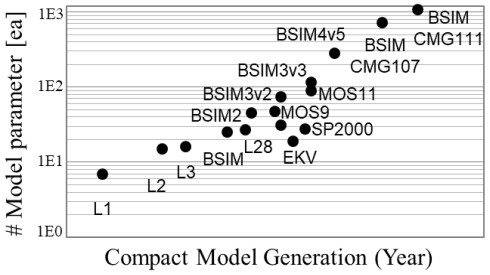

Fig. 1. The number of standard compact model's fitting parameters. It is rapidly increasing because of modern device's secondary effects and non-monotonic characteristics

III. PHYSICS-AUGMENTED NEURAL COMPACT MODEL

Artificial intelligence is providing effective alternatives to various Electronic Design Automation (EDA) technologies, and outperformed the existing competitors. Therefore, combining these data science techniques (artificial neural networks) and scientific theories (device physics) are able to have the best of both worlds [2-4] for SPICE application as shown in Fig. 2.

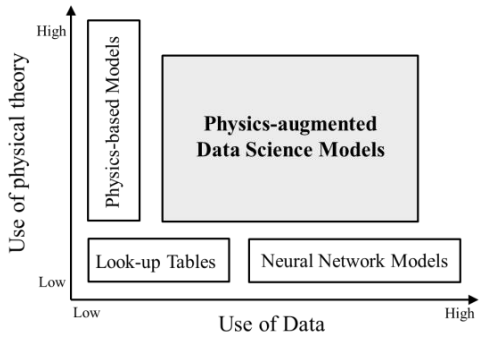

Fig. 2. Physics-augmented neural network model for circuit simulations

This modeling framework includes two main steps, 1) data preprocessing according to various physical origins and machine learning training with each physical constraint loss, 2) model validation (accuracy test), model compression (model scaling for computational efficiency) and charge-conservative model implementations as shown in Fig. 3. First of all, electrical measurement data of various circuit elements could be componentized using the scientific domain knowledges and physical origins as shown in Fig. 4. To briefly demonstrate the component-wise modeling flow, we select a modern MOSFET device as an example as shown in Fig. 3(b). Drain current (drift-diffusion transport), gate current (tunneling), and body current (gate-induced drain/source leakage) are main target components according to design instance (W/L/T), bias (V_{gs}/V_{ds}/V_{bs}) and technology (T_{ox}/WF) parameters.

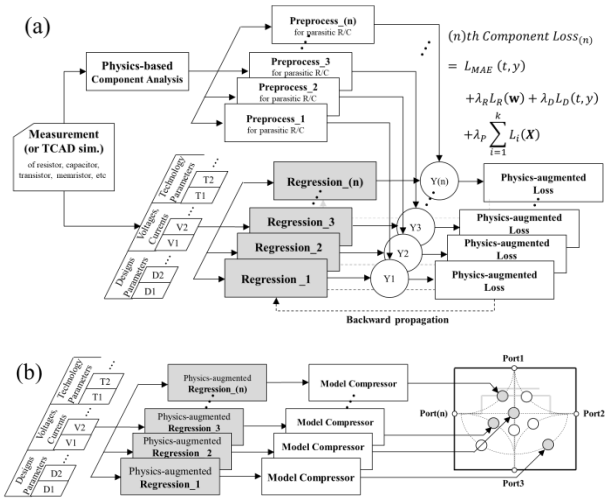

Fig. 3. Physis-augmented neural compact modeling framework. (a) Component-wise data prepocessing and physics-augemented machine learning training, (b) Component-wise model validation and compression for accuracy and computational efficiency in SPICE applications

Elements (e.g.)		Components examples
Capacitor	Passive	Area cap + Fringing cap
	MOS	Intrinsic charge cap + Extrinsic cap
Transistor	Bipolar	Hole current + Electron current
	Unipolar	Transport + body + tunneling currnent

Fig. 4. An example of circuit element's electrical and physical components

Parameter	Description
X	Number of input features in the neural networks
Y	Number of physical components in the equvalent circuit
L	Number of hidden layer in the neural networks
Nl	Number of neurons of lth layer
λp	Physics augmentation loss coefficient
λa	Derivative loss coefficient for analog models
λd	Weight decay parameter for digital models
wji,l	Weight from ith neuron of the (l-1)th layer to the jth neuron of the lth layer
bi	Bias of the ith neuron of the lth layer

Fig. 5. Model parameters to demonstrate the component-wise neural network modeling flow in this work

These component data pass through each hypothesis space to find optimal model parameters, and in this demonstration, fully connected neural networks are used to capture each target function as shown in Fig. 5. The trained models are implemented as an equivalent-circuit form to meet the charge conservation raw, and the SPICE result shows a good agreement between the model inferences and corresponding targets as shown in Fig. 6.

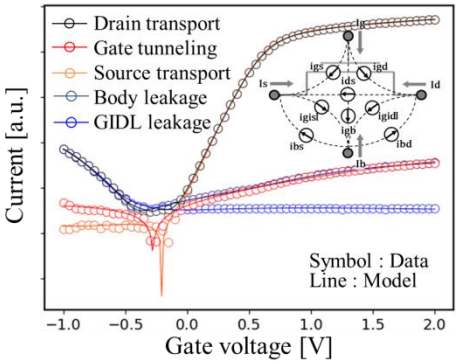

Fig. 6. Model inference results which meet the charge conservation raw

A. *Physics-augmented model*

In this paper, physics-augmented neural network model is introduced to provide an interpretable and physically consistent circuit simulations. For example, MOSFET's I-V transfer characteristics has a distinct temperature dependency (i.e. trend reversion by an zero temperature coefficient point) because dominant transport mechanisms are changed according to the bias and electric field conditions as shown in Fig. 7.

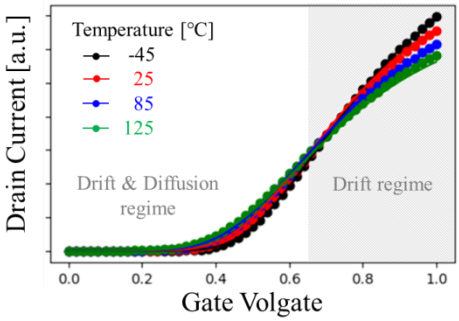

Fig. 7. MOSFET's I-V transfer curve which has an reversed temperature dependency according to the bias conditions

However, measurements with various temperatures are time-consuming and the data is costly. Therefore, most model parameter extraction procedures are based on few temperature sweeps. Fig. 8 shows the ordinary neural network modeling results which are trained using these costly data at -25 and 125 degree Celsius only, and it has physically wrong inferences around the unseen data. These 'black-box' neural network models are prone to small data set and unintended noises which is most common at SPICE modeling environments.

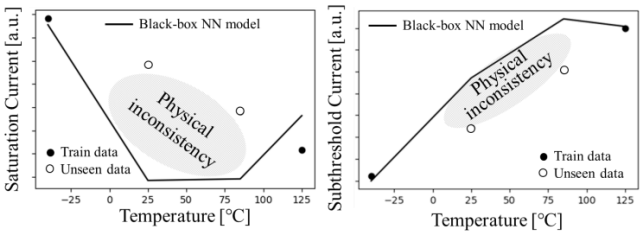

Fig. 8. Neural network model training results (left: saturation currents, right: subthreshold currents) based on cool and hot temperatures only. These black-box models are vulnerable to the unseen data, and it has physically reversed temperature coefficients around 25~85 degrees Celsius

Therefore, we need to make the most of the physical relationships between sample data (transport currents) and input features (bias voltages, temperatures) as follows.

Temperature (T) is closely related to the carrier mobility (μ) and Fermi potential (Φ_F) in the saturation and subthreshold regions respectively as below,

$$log(\mu) = \tau_c \cdot log(T)^{-\frac{3}{2}} \quad (if\ V \gg V_{th}) \tag{1}$$

$$\phi_F = \frac{kT}{q} \cdot ln\left(\frac{N_a}{n_i}\right) \quad (if\ V \ll V_{th}) \tag{2}$$

τ_c is an impurity scattering related constant, N_a is substrate doping concentration, n_i is intrinsic carrier concentration and k is Boltzmann constant. The carrier mobility (μ) in the saturation region is closely related to drift

currents and Fermi potential (Φ_F) in subthreshold region also has a physics-based relationship with diffusion currents as shown in Fig. 9-10.

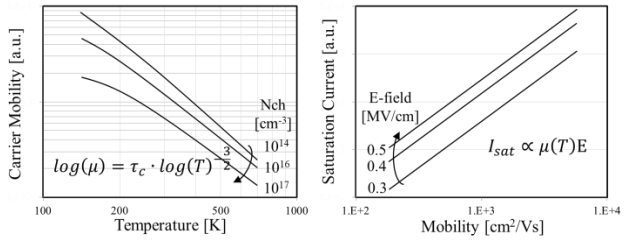

Fig. 9. The temperature effects on the carrier mobility (left) and saturation currents (right)

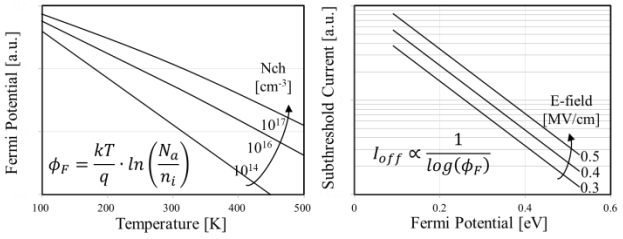

Fig. 10. The temperature effects on the Fermi potentials (left) and subthreshold currents (right)

Therefore, the carrier mobility and Fermi potential at two different temperatures ($T_1 < T_2$) on the same device size (W, L) are related to each other in the following manner,

$$\mu[T_2, W, L, V_{dH}] - \mu[T_1, W, L, V_{dH}] \le 0 \quad (V_{dH} \gg V_{th}) \quad (3)$$

$$\phi_F[T_2, W, L, V_{dL}] - \phi_F[T_1, W, L, V_{dL}] \le 0 \ (V_{dL} \ll V_{th}) \quad (4)$$

It also means we are able to compute the differences in mobility and Fermi potential of a neural network model, based on any pair of consecutive temperatures, T_i and T_{i+1} ($T_i<T_{i+1}$) in a device size (W, L) as below,

$$\Delta\mu = \mu[T_{i+1}, W, L, V_{dH}] - \mu[T_i, W, L, V_{dH}] \quad (5)$$

$$\Delta\phi_F = -\phi_F[T_{i+1}, W, L, V_{dL}] + \phi_F[T_i, W, L, V_{dL}] \quad (6)$$

A positive value of these differences are considered as a physical consistency based on the theory equation (1-2) on the temperature and device size. Therefore, we can also construct physics-augmented loss functions to guide the neural networks toward a negative mean of all consecutive temperature pairs (T_i, T_{i+1}) using the Rectified Linear Unit activation function (ReLU) [5].

$$Physics\ augmented\ loss\ of\ drift\ components = \frac{\lambda_P}{n_w n_l (n_t - 1)} \sum_{w=1}^{n_w} \sum_{l=1}^{n_l} \sum_{i=1}^{n_t - 1} ReLU(\Delta\mu[T_i, W, L, V_{dH}]) \quad (7)$$

$$Physics\ augmented\ loss\ of\ diffusion\ components = \frac{\lambda_P}{n_w n_l (n_t - 1)} \sum_{w=1}^{n_w} \sum_{l=1}^{n_l} \sum_{i=1}^{n_t - 1} ReLU(\Delta\phi_F[T_i, W, L, V_{dL}]) \quad (8)$$

A neural network model is trained using same limited data (-25 and 125℃ only) , however in this case, the device physics-informed losses are also used to search the optimum solution. Fig. 11 shows the physically correct inference results at both drift and diffusion regions, and the fitting accuracies are greatly improved by introducing these scientific theory-guided learning algorithms.

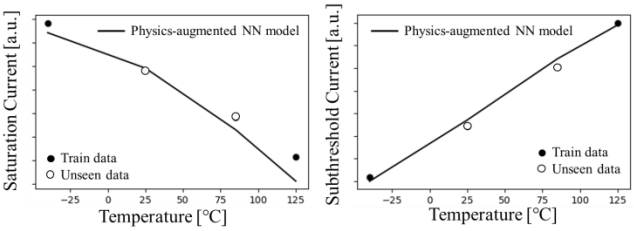

Fig. 11. Neural network model training results (left: saturation currents, right: subthreshold currents) based on cool and hot temperatures. In this case, physics augmented losses are applied, and it has physically correct inferences in the unseen data regions

B. *Circuit-friendly model*

Semiconductor data which pass through the neural networks are inherently vulnerable to inevitable measurement noises and electrical fluctuations. In addition, robust compact models require to satisfy the high order derivatives and symmetry characteristics for various digital and analog circuit applications. A transistor's on-off ratio is one of the most important performance factors in digital circuit analysis, and the accurate leakage current modeling is critical to PPA assessments. Therefore, the weight decay techniques are essential to these digital models as shown in Fig. 12, and the regularization strengths need to be properly adjusted according to the data qualities as below,

$$Regularization\ loss = \frac{\lambda_R}{2} \sum_{i=0}^{m} {W_i}^2 \quad (9)$$

λ_R is regularization coefficient, and W_i stands for weight parameters of a regression model.

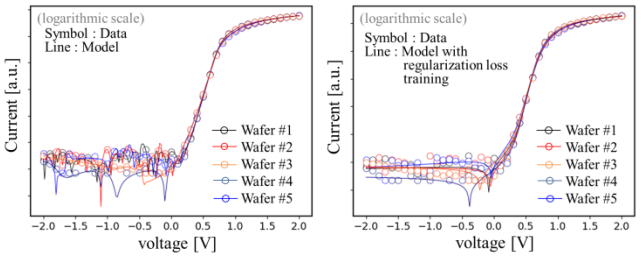

Fig. 12. Semiconductor measurement data which include random noises and elecrical fluctuations. Practical neural models require to use effective regularization losses (right) for avoiding the overfitting to noises (left)

The relationships between terminal current and conductance or terminal charge and capacitance are also important features in device modeling. An accurate inference of transconductances from current model or capacitances from charge model are all dependent on derivative considerations. These derivative characteristics are critical, not only for analog circuit analysis but also for iterative solving algorithm of SPICE simulators. Therefore, high order derivatives need to be introduced in the loss calculation as shown in Fig. 13.

$$Deriv\ loss = \frac{\lambda_D}{2} \sum_{n=1}^{\delta} \sum_{i=0}^{m} \left(\frac{d^{(n)} t_i}{dt^{(n)}} - \frac{d^{(n)} y_i}{dy^{(n)}} \right)^2 \quad (10)$$

λ_D is derivative coefficient. t_i and t_i are true and estimation value, respectively.

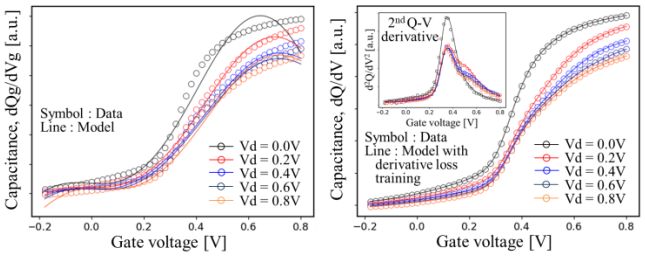

Fig. 13. An ordinaly neural charge model's inaccurate inference results of gate capacitance (left). Higher order derivtive loss is able to dramatically impoves the accuracy of derivative characteristics.

IV. Model Validations

A. Scalablility in DRAM applicatoins

This modeling framework is applied to a DRAM technology. Fig. 14 shows a highly scalable and accurate fitting ability of neural compact model, which is contrast to a standard compact model (scalable, inaccurate) and empirical binning approach (non-scalable, accurate).

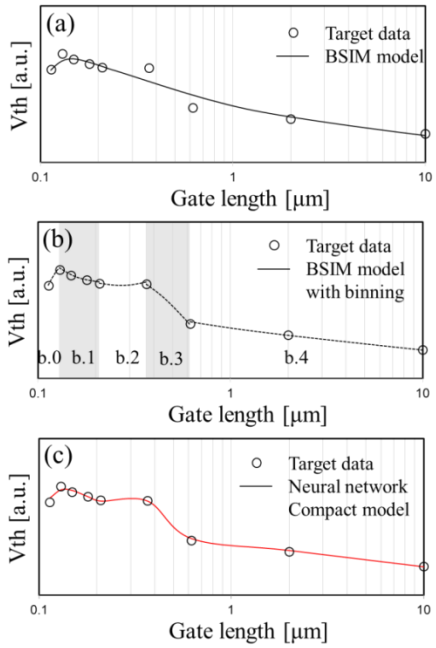

Fig. 14. Model scalability comparison between standard compact model, empirical binnging and neural compact model.

B. Technology-independence in Logic applications

It is also applied to various state-of-the-art emerging devices [6-10] as shown in Fig. 15. These new technologies are not supported by existing standard compact models. Fig. 16 shows a highly general fitting ability to cross-sectional size dependencies such as multiple humps and peaks in C-V and steep switching characteristics in I-V curves.

	Emerging Device (a)	Emerging Device (b)	Emerging Device (c)	Emerging Device (d)
Technology	NanoWire	Nano Wire	NCFET	Tunneling FET
Structure	Gate-All-Around	Gate-All-Around	Tri-gate	Vertical
Ch. Material	Si	InGaAs	Si	Si
EOT	1nm	1nm	-	0.6nm
Dielectrics	SiO2	SiO2	HfO/TaN	SiO2
Device Features	1-D cross sectional QME in Si	1-D cross sectional QME in III-V	Low SS < 60mV/Dec	Low On/Off ratio <1e-9

Fig. 15. Various state-of-the-art emerging devices for early stage PPA and DTCO benchmarks in this work

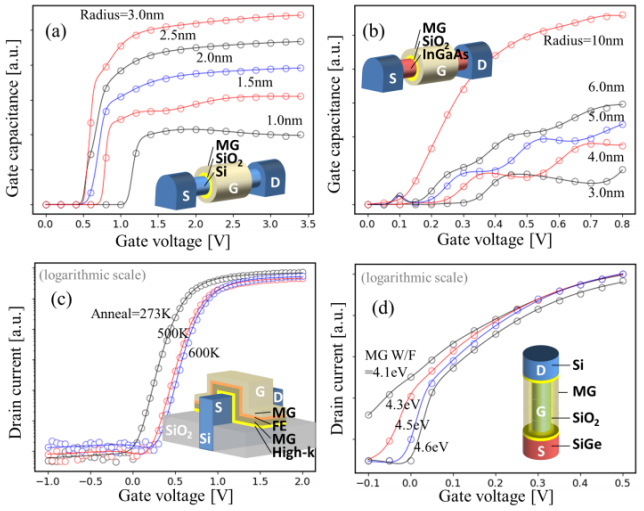

Fig. 16. Neural SPICE model results of various emerging devices. (a) Ultra narrow silicon nanowire, (b) Ultra narrow InGaAs nanowire, (c) Tri-gate negative capacitance transistor and (d) Vertical tunneling transistor

V. Conclusion

A series of neural network modeling algorithms and physics informed machine learning techniques are proposed. Semiconductor data passes through each neural networks according to its physical origin, and parameters are automatically updated in backpropagation with each physics-based constraint loss. This physics-augmented neural compact model has the best of two worlds, neural network model's fitting accuracy and theory-based model's physical consistency. It is also applied to recent DRAM and promising Logic technologies, and evaluated in terms of its scalability and technology independency. This new approach is expected to be an effective alternative for early stage PPA and DTCO activities for new emerging devices.

References

[1] Yohan Kim et al., "The efficient DTCO Compact Modeling Solutions to Improve MHC and Reduce TAT", SISPAD, September 2018.

[2] Anuj Karpatne et al.,"Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data", IEEE Transs on Knowledge and Data Engineering, vol. 29, pp. 2318-2331, June 2017.

[3] Anuj Karpatne et al., "Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling", arXiv preprint arXiv:1710.11431, 2017.

[4] Xiaowei Jia et al., "Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles", arXiv preprint arXiv:2001.11086, 2020.

[5] Xavier Glorot et al., "Deep sparse rectifier nueral networks", International Conference on Artificial Intelligence and Statistics, vol. 15, pp. 315-323, 2011.

[6] Avirup Dasgupta et al., "Unified Compact Model for Nanowire Transistors Including Quantum Effects and Quasi-Ballistic Transport" IEEE Trans on electron devices, vol. 64, no. 4. pp. 1837-1845, April 2017.A

[7] Avirup Dasgupta et al., "Compact Modeling of Cross-Sectional Scaling in Gate-All-Aroud FETs: 3-D to 1-D Transition", IEEE Trans on electron devices, vol. 65, no. 3, pp.1094-1100, March 2018.

[8] Kai-Shin Li et al., "Sub-60mV-Swing Negative-Capacitance FinFET without Hysteresis", IEDM, December 2015.

[9] S. Khandelwal et al., "Circuit Performance Analysis of Negative Capacitance FinFETs", IEEE Symposium on VLSI Technology, June 2016.

[10] Eunah Ko et al., "Vertical Tunneling FET: Design Optimization With Triple Metal-Gate Layers", IEEE Trans on Electron Devices, vol. 63, December 2016.