Using Long Short-Term Memory (LSTM) Network to Predict Negative-Bias Temperature Instability

Fanus Arefaine
Dept. of Computer Engineering
San Jose State University
San Jose, USA
fanus.arefaine@sjsu.edu

Meng Duan
Synopsys Northern Europe, Ltd.
Glasgow, United Kingdom
meng.duan@synopsys.com

Ravi Tiwari
Dept. of Electrical Engineering
Indian Institute of Tech., Bombay
Mumbai, India
ravi.fermi@gmail.com

Aadit Kapoor
Dept. of Computer Science
San Jose State University
San Jose, USA
aadit.kapoor@sjsu.edu

Lee Smith
Synopsys, Inc.
Mountain View, USA
Lee.Smith@synopsys.com

Souvik Mahapatra
Dept. of Electrical Engineering
Indian Institute of Tech., Bombay
Mumbai, India
souvik@ee.iitb.ac.in

Hiu Yung Wong*
Dept. of Electrical Engineering
San Jose State University
San Jose, USA
hiuyung.wong@sjsu.edu

***Abstract*— In this paper, Long Short-Term Memory (LSTM) is used to predict transistor degradation due to Negative-Bias Temperature Instability (NBTI). The LSTM is trained by Technology Computer-Aided Design (TCAD) generated NBTI data and then used to predict the future degradation based on the future stress pattern (i.e. the future gate voltage sequence). It is also used to predict the degradation due to other random stress patterns at different frequencies. It is found that the LSTM trained by NBTI data due to random gate pulses at 100MHz clock frequency can 1) predict the NBTI due to other random gate pulses, 2) predict the NBTI up to 2 times longer time than it is trained for, and 3) predict the NBTI of 10 times higher and lower clock frequencies. Moreover, it can capture the Transient Trap Occupancy Model (TTOM) and Activated Barrier Double Well Thermionic (ABDWT) models well. It is shown that the framework works for both 2D and 3D simulations and, thus, can save a substantial amount of TCAD simulation time.**

Keywords—Degradation, Long Short-Term Memory (LSTM), Negative-Bias Temperature Instability (NBTI), Reliability, Technology Computer-Aided Design (TCAD)

I. Introduction

Negative-Bias Temperature Instability (NBTI) is an important degradation mechanism that has received a lot of attention in the modeling community [1]-[5]. Among them, Reaction-Diffusion (R-D) model is one of the most practical and promising ones that has been demonstrated in 3D FinFET NBTI modeling [2][3]. The R-D model is made complete by including the Transient Trap Occupancy Model (TTOM) [1] and the Activated Barrier Double Well Thermionic (ABDWT) model [6] to account for trap occupation and hole trapping/emission, respectively. However, AC transient simulation of NBTI under MHz-GHz gate voltage sequence is computationally intensive.

In this paper, Long-Short-Term Memory (LSTM) [8], a type of recurrent neural network (RNN), is used. RNN has been shown to be promising in modeling transient circuit simulations [9][10]. Since LSTM can avoid the vanishing gradient problem, it is chosen in this study. It is assumed that some degradation data has been created through TCAD simulation under random gate pulses for a certain period of time. R-D model with TTOM and ABDWT is used for realistic and accurate simulations. The data are expressed as the average interfacial trap oxide (N_{it}) and the average hole trap charge (ABDWT charge) as a function of time and gate voltage sequence. An LSTM machine is trained using the TCAD data. The abilities of the trained machine to predict the degradation due to 1) unseen random gate voltage sequence, 2) future gate voltage sequence, and 3) unseen gate voltage sequence at different frequencies are studied. Both 2D and 3D TCAD simulation data are presented.

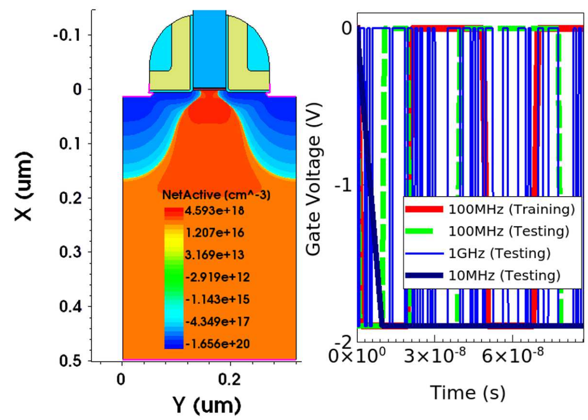

Figure 1: Left: The 2D structure used for TCAD simulation. Right: Random gate pulses used in the simulation. Only limited time is shown for clarity. Each curve is discretized by 1/10th of the period for machine learning. V_G is between 0V and -1.9V.

II. TCAD Simulation and Data Preparation

TCAD Sentaurus is used for 2D PMOS creation and simulation [7]. Besides the Poisson equation, and electron/hole continuity equations, hydrogen atom, and molecule diffusion equations are also solved. The density gradient equation is included to account for the quantum confinement effect. Multi-State-Configuration (MSC) in Sentaurus Device is used to model the changing of the interface states (Si-H, X-H, Si^+, Si, etc.) in the R-D model. TTOM and ABDWT are turned on. Fig. 1 shows the structure created and the random gate pulses used in the TCAD simulations with V_D = 0V and ambient temperature of 398K.

*Corresponding Author: hiuyung.wong@sjsu.edu

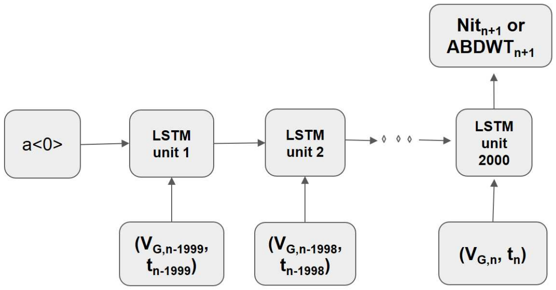

Figure 2: LSTM used in this study. Two machines are built, one for predicting the N_{it} and one for predicting the ABDWT charge.

The training data, namely the interface trap density (N_{it}) and the ABDWT charge (oxide trapped charge), is generated by using a 100MHz clock (i.e. period = 10ns) with random gate pulses and the simulation results are discretized every 1ns. Three testing data with different random gate pulses, namely 10MHz, 100MHz, and 1GHz, are generated (Fig. 1). Each of them is discretized at 1/10$^{\text{th}}$ of the corresponding period (e.g. 1GHz is discretized at 0.1ns interval). The simulation is performed on Intel Xeon Gold 6254 3.1GHz CPU. Every 1000 data points take about 5.4 hours.

III. LSTM Training

5000 points, i.e. 5µs of data, from the training data are used to train the LSTM (Fig. 2). The LSTM is optimized and it is found that 2000-LSTM-unit performs the best. *tanh* is used for activation which provides a significant speedup over *ReLU* during training probably due to its less expensive gradient computation. Many-to-one LSTM architecture is used and the N_{it} or ABDTW charge at a certain time is determined by the previous 2000 gate voltages (V_G) and times (t). Two LSTMs are trained, one for predicting N_{it} and one for predicting the ABDWT charge. They are trained for 883 and 251 epochs, respectively. The time required to train each machine is less than 90 minutes on an NVIDIA Quadro GPU.

IV. LSTM Performance

The trained LSTM is then used to predict the NBTI degradation process of the transistor under three unseen testing pulses.

Prediction of new random gate voltage sequence and longer gate voltage sequence:

Fig. 3 shows the prediction of unseen new random gate pulses at the clock frequency of 100MHz for 10µs. The LSTM can predict the change of N_{it} and ABDWT charge well over 10µs even it has not seen the pulses and it was only trained for 5µs. In particular, it can capture the fast recovery due to unoccupied traps (TTOM). The prediction process takes less than 1 minute which represents a significant speedup compared to the time required to simulation 10µs of new random gate pulses which will take about 54 hours.

Prediction of new random gate voltage sequence of different frequencies:

The same models trained by the 5µs 100MHz data are then applied to predict the NBTI degradation of the same device at different frequencies (10MHz and 1GHz) for the same number of data points, i.e. 5000. Therefore, 50µs and 500ns of degradation are predicted for 10MHz and 1GHz data, respectively. Fig. 4 and Fig. 5 show the Seaborn joint plots of the LSTM prediction and TCAD simulation. Despite the model being used to predict the degradation at different timescale and

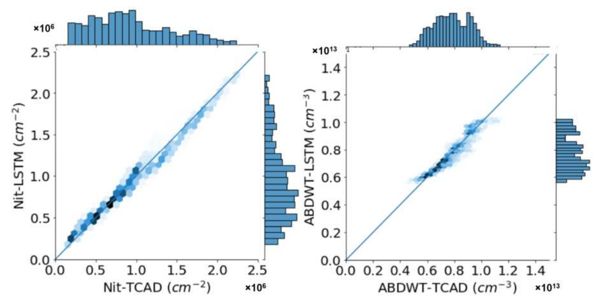

Figure 4: Seaborn joint plots of TCAD and LSTM NBTI prediction for clock frequency of 10MHz using the model trained by 100MHz. Left: N_{it}. Right: ABDWT charge.

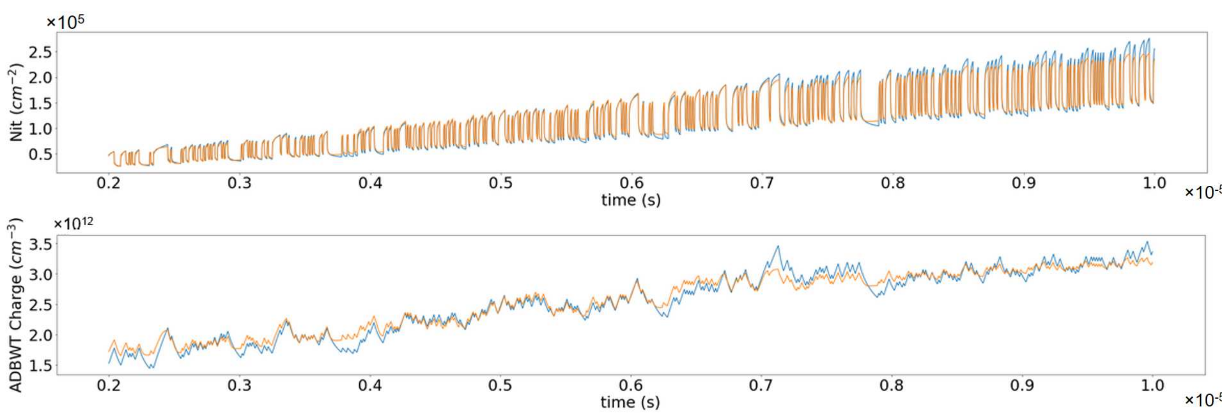

Figure 3: Comparison between LSTM prediction and TCAD simulation for N_{it} (top) and ABDWT charge (bottom) for 100MHz testing data up to 10µs generated by unseen random gate voltage pulses for the 2D structure in Fig. 1. LSTM was trained by another set of 5µs 100MHz data. Orange: LSTM. Blue: TCAD.

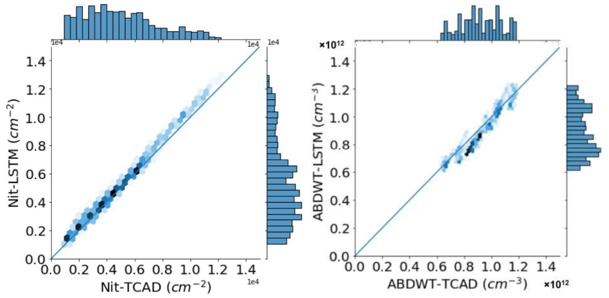

Figure 5: Seaborn joint plots of TCAD and LSTM NBTI prediction for clock frequency of 1GHz using the model trained by 100MHz. Left: N_{it}. Right: ABDWT charge.

In all the studies conducted, LSTM trained by T_I amount of data usually can predict the degradation of the next T_I amount of time fairly well. This is shown in Fig. 3 and Fig. 6. In Fig. 6, the machine is trained by the first 10µs of data with 2000-LSTM units and it is used to predict the next 20µs (i.e. additionally $2T_I$). While it can predict the 10µs-to-20µs degradation well, the prediction of the 20µs-to-30µs one becomes worse. It can be seen that between 20µs to 30µs, the machine can predict the increase and the trend of N_{it} and ABDWT charge and also can track the change of gate voltages (i.e. degradation increases when the gate pulse is negative and recovers when the gate pulse is 0V). However, it cannot predict the amplitude of the fluctuation of these quantities. The amplitudes indeed stay almost constant after 30µs. Moreover, after 40µs (not shown), it can no longer predict the increasing trend of the degradation.

the N_{it} or ABDWT charge span a range of 100 times, the predictions are good. The R^2 scores of 100MHz and 1GHz are larger than 0.98 and 0.94, respectively.

V. Study of LSTM Prediction Capability

To further understand the role of the number of LSTM units and the capability and limitation of the LSTM machine, different LSTM machines are trained by 100MHz random gate voltages for different times, T_I, namely 5µs, 10µs, and 20µs with different numbers of LSTM units (2000, 4000, and 8000). Their respective ability to predict the degradation in the next T_I amount of time (e.g. next 20µs for the 20µs trained machine) is studied by comparing the R^2 score. However, no trend can be concluded probably due to the competition between overfitting and length of historical gate voltage data. With more LSTM units, it is easier to have overfitting. On the other hand, more LSTM units can store a longer history of gate pulses and can help to predict future degradation better. Therefore, a careful choice of the number of LSTM units is important. For example, 2000-unit is found to be the best for the 5µs and 10µs trained machines while 8000-unit is found to be the best for the 20µs trained machine. All these machines are trained for 1000 epochs.

VI. 3D Simulation Data Prediction

The same approach is applied to 3D TCAD data. Since 3D simulation is much slower than 2D, this methodology is expected to save a more substantial amount of simulation time. A 3D FinFET is constructed with a Fin width of 8nm, a Fin height of 42nm, and a gate length of 20nm (Fig. 7). It has 1.7nm HfO_2 and 0.82nm interfacial oxide as the gate insulator. To reduce the simulation time, only half of the structure is simulated. Moreover, a long metal is added to mimic the long diffusion path of chemical species such as H_2 across the die. The structure has ~140,000 mesh points. The same set of equations and models are solved as in the 2D cases. 4 CPU cores are used in the simulation. It takes about 20 days to complete the simulation of 10µs of 100MHz random gate voltages. The first 5µs is used for training with 2000 LSTM units and the trained machine is then used to predict the degradation of the last 5µs. V_G is between 0V and -1.4V and the ambient temperature is 398K. The machine is trained for 2000 epochs.

Fig. 8 shows the prediction results. The trained LSTM machine can predict the degradation well with R^2 scores of 0.96 and 0.92 for N_{it} and ABDWT charge, respectively. Note that the degradation is more severe in FinFET than in the 2D structure.

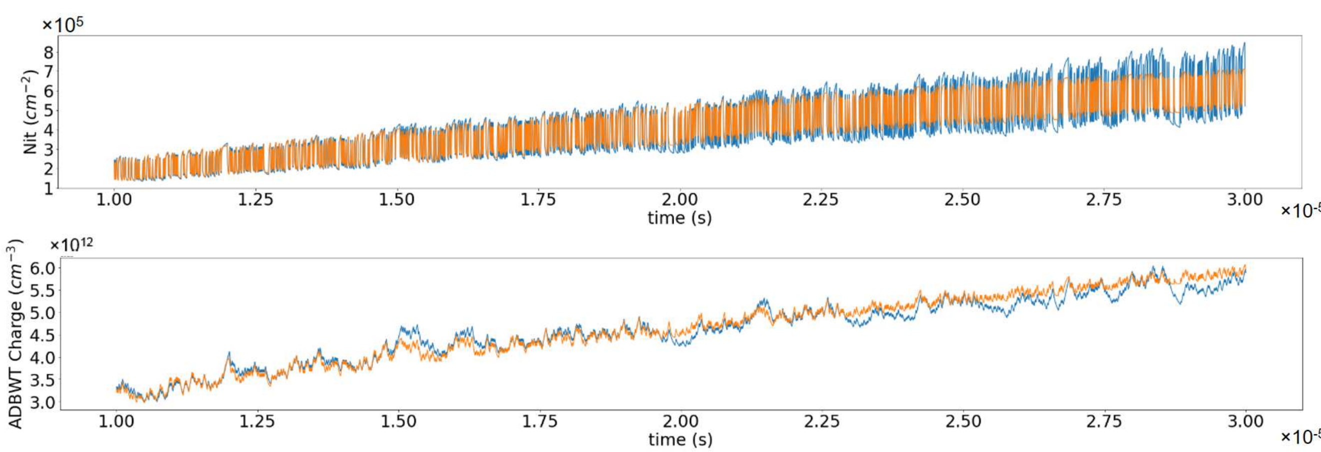

Figure 6: Comparison between LSTM (trained by T_I=10µs data) prediction and TCAD simulation for N_{it} (top) and ABDWT charge (bottom) for 100MHz testing data. For clarity, only the prediction between 10µs and 30µs is shown. Orange: LSTM. Blue: TCAD.

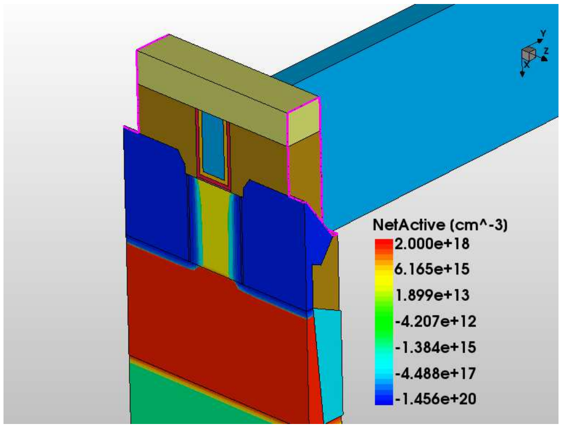

Figure 7: 3D FinFET used for NBTI simulation. Half structure is used to reduce the simulation time. A long metal gate (blue) is used to allow chemical species, such as H_2, to diffuse in a large enough domain to establish realistic boundary conditions.

Moreover, even though it also has a less steep slope in the region of study, the LSTM framework still works well.

VII. Conclusions

In this paper, the LSTM model is applied to predict the NBTI degradation of transistors. The model is trained by TCAD data generated with a sequence of random gate pulses. With proper training, the model is able to predict the degradation under new random gate pulses up to two times longer time. It also can predict the degradation due to gate pulses with other frequencies (10 times higher and 10 times lower). It can capture not just the R-D model but also the TTOM and ABDWT models well. It is also shown that this can be used to predict 3D TCAD degradation data and 10 days of simulation time can be saved in the case demonstrated. With proper simplification, this model may be used as a compact model for circuit simulations.

Acknowledgment

This research is partially supported by the Level-Up Grant, Office of Research, San Jose State University (SJSU). Part of the machine learning was performed using the SJSU College of Engineering (CoE) High Performance Computing (HPC) system.

References

[1] N. Parihar et al, “BTI Analysis Tool—Modeling of NBTI DC, AC Stress and Recovery Time Kinetics, Nitrogen Impact, and EOL Estimation”, IEEE Transactions on Electron Devices, Vol. 65, No. 2, 2018.

[2] R. Tiwari et al., "A 3-D TCAD Framework for NBTI—Part I: Implementation Details and FinFET Channel Material Impact," in IEEE Transactions on Electron Devices, vol. 66, no. 5, pp. 2086-2092, May 2019, doi: 10.1109/TED.2019.2906339

[3] R. Tiwari et al., “A 3-D TCAD Framework for NBTI, Part-II: Impact of Mechanical Strain, Quantum Effects and FinFET Dimension Scaling, in IEEE Transactions on Electron Devices, vol. 66, no. 5, pp. 2093-2099, May 2019.

[4] T. Grasser et al., “A Two-Stage Model for Negative Bias Temperature Instability,” in IEEE International Reliability Physics Symposium (IRPS), Montréal, Québec, Canada, pp. 33–44, April 2009.

[5] V. Huard, “Two independent components modeling for negative bias temperature instability,” in Proc. Int. Rel. Phys. Symp., 2010, pp. 33–42.Arute, F., Arya, K., Babbush, R. et al. Quantum supremacy using a programmable superconducting processor. Nature 574, 505–510 (2019). https://doi.org/10.1038/s41586-019-1666-5A.

[6] N. Choudhury et al., “A Model for Hole Trapping-Detrapping Kinetics During NBTI in p-Channel FETs,” IEEE Electron Devices Technology and Manufacturing (EDTM) Conference, March 2020.

[7] Sentaurus^{TM} Device User Guide, Synopsys Inc., Mountain View, CA, USA, 2020.

[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation,9(8),1735–1780, 1997.

[9] T. Nguyen, T. Lu, K. Wu, and J. Schutt-Aine, “Fast Transientsimulation of High-Speed Channels using Recurrent NeuralNetwork,” 2019. [Online]. Available: https://arxiv.org/abs/1902.02627

[10] Z. Chen, M. Raginsky and E. Rosenbaum, "Verilog-A compatible recurrent neural network model for transient circuit simulation," 2017 IEEE 26th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS), 2017, pp. 1-3, doi: 10.1109/EPEPS.2017.8329743.

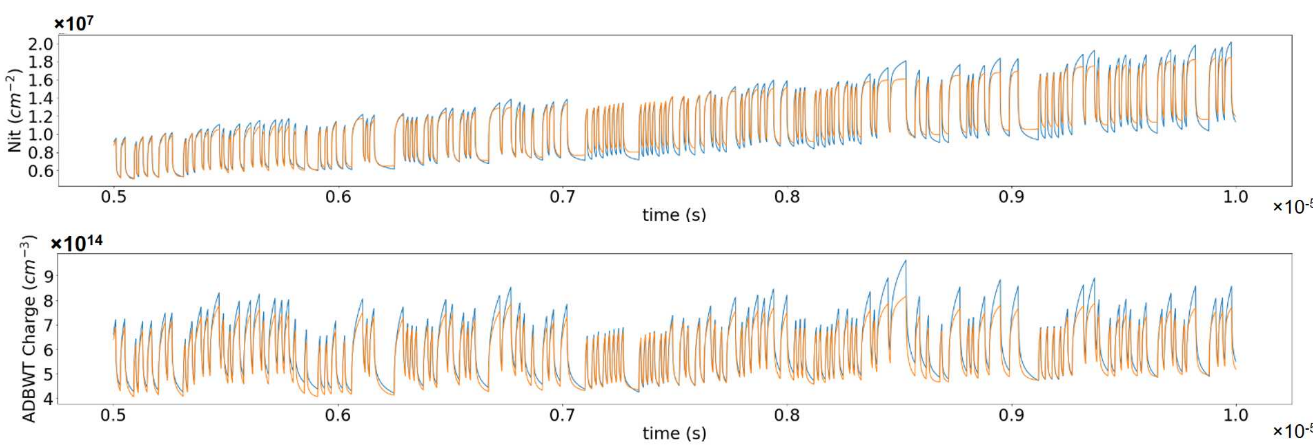

Figure 8: Comparison between LSTM prediction and TCAD simulation for N_{it} (top) and ABDWT charge (bottom) for 100MHz testing data between 5µs and 10µs generated by unseen random gate voltage pulses for 3D FinFET. The LSTM was trained by the first 5µs data (not shown). Orange: LSTM. Blue: TCAD.