Symposium pour l'électronique & le numérique durables
Le 12 décembre 2024, Grenoble

LSTM surrogate model for SiGe HBT optimization

Caron Grégoire
Laboratoire Jean Kuntzmann
Saint-Martin-d'Hères, France
g.caron1789@gmail.com

Juditsky Anatoli
Laboratoire Jean Kuntzmann
Saint-Martin-d'Hères, France
anatoli.iouditski@univ-grenoble-alpes.fr

Guitard Nicolas
STMicroelectronics
Crolles, France
nicolas.guitard@st.com

Céli Didier
STMicroelectronics
Crolles, France
didier.celi@st.com

Abstract—Silicon-germanium heterojunction bipolar transistors are becoming increasingly miniaturized, making design difficult and requiring resource-intensive fabrication. While simulation tools aid optimization, they are computationally intensive. This paper proposes to build a calculation friendly database to train a reduced mathematical model of the tool and eventually allow for more ambitious applications free from expensive simulations.

Keywords—computational resources, heterojunction bipolar transistor, TCAD, neural networks

I. Introduction

Silicon-germanium (SiGe) heterojunction bipolar transistors (HBTs) are used in various power and high-frequency applications. To improve their performance and reduce power consumption, they are constantly being miniaturised. This makes their design complex and requires the fabrication of numerous devices on silicon wafers - a lengthy process that consumes water, electricity and chemicals - before converging on an optimized device. To partially overcome these drawbacks, Technology Computer-Aided Design (TCAD) is used to simulate the electrical characteristics of transistors. However, this requires numerous costly numerical calculations. In this paper we present a technique based on the use of "reduced" mathematical models which, once built, replace TCAD. In a first part, we present a comprehensive methodology used to build a database. Then, the "surrogate" models are built using recurrent neural networks. Finally, we present two important applications to illustrate how this technique eliminates the need for a large number of costly simulations.

II. Database Generation

First, we build a database of pairs of "1D doping profiles of transistors; electrical characteristics simulated by Sentaurus SDevice". Profiles are described analytically using a representation $N(x) = N_0 \exp(-a^{-1}|x-x_0|^b)$ [1] with parameters $N_{0,}$ a, b and x_0 for each part (emitter, base, collector and germanium). The chosen characteristics are the capacitances, collector current, sheet base resistance and transit frequency.

To limit the size of the database - and therefore the number of simulations required to generate it - we focus on profiles that meet two conditions. First, they must be realistic, i.e. meet basic criteria on the doping level, thickness and position of each part. To achieve this, we use Monte Carlo rejection sampling accelerated by a binary classification method (Support Vector Machine) [2]. Second, they must be technologically interesting, i.e. have high transit frequencies f_{Ti}. We use adaptive importance sampling with Gaussian density estimation [2] to gradually target these transistors of interest in the parameter space, as shown in Fig.1. These methods allow to limit the size of the database to 60,000 records "profile – characteristics". This remains a large resource-investment, but it is a one-time cost: once the database is built and the surrogate models are trained, one can design 1D profiles without using TCAD.

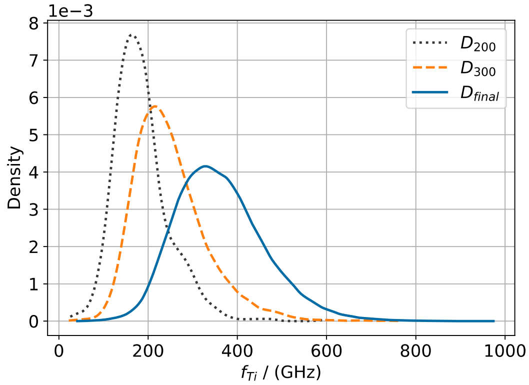

Fig.1: Adaptive importance sampling focused on high f_{Ti}

III. Surrogate Model Training

We build individual models to predict each output characteristic from input profile parameters. The characteristics are regularly sampled in electrical bias V_{BE}. Using Python library Tensorflow Keras, neural networks (NNs) are trained for this regression task, in which the weights linking layers of units are optimized by backpropagation and stochastic gradient descent. The most common NNs are called Feedforward Neural Networks (FNNs). However, the values of adjacent points are inherently correlated in sampled characteristics. Therefore, Long Short-Term Memory (LSTM) networks [3] are chosen to account for this, which are a type of NNs specifically designed to handle long sequential data. We adopt a decoder-inspired architecture [4], shown in Fig.2. It involves considering the profile as a latent representation of the electrical characteristics to decode.

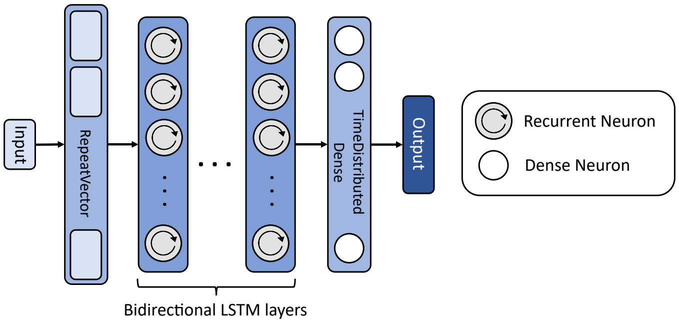

Fig.2: LSTM decoder architecture

We assess the accuracy on test data not seen during training, using the Relative Absolute Deviation as a metric. Using LSTM decoders leads to important gains (35% to 60%) in prediction accuracy over FNNs, with less than 2% error. As importantly, all characteristics are retrieved almost instantly instead of tens of seconds when using SDevice.

IV. Applications

Profile improvement. We present a method to represent the sensitivity of output characteristics y_j to input parameters x_i. The models allow for an accurate computation of gradients $\partial y_j / \partial x_i$ at each bias point V_{BE}, averaged on thousands of transistors, which is intractable with TCAD alone. For example, Fig.3 illustrates the sensitivity of f_{Ti} to inputs x_i, showing that at high V_{BE}, collector doping $N_{0,C}$ is very positively correlated with f_{Ti}. This validates the model's consistency with transistor physics and highlights improvement opportunities.

Design Space Exploration. Another essential application involves searching for efficient transistors, by exploring the entire space of profile parameters. Hundreds of thousands of synthetic profiles are generated. Their characteristics are then predicted using the models much faster than with the simulation tools, saving weeks of computation and energy consumption.

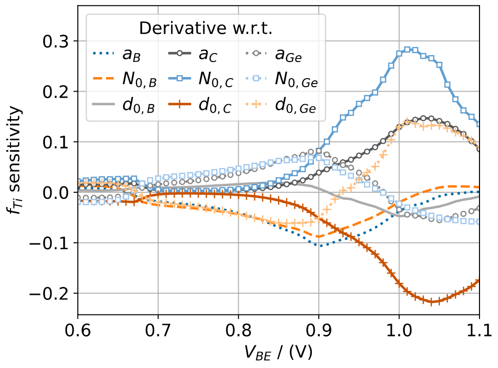

Fig.3: Evolution of f_{Ti} sensitivity to important input parameters with bias

Figure 4 presents a map of the average f_{Ti} derived from Principal Component Analysis (PCA), highlighting regions of profiles associated with higher f_{Ti} values. These areas can be targeted while imposing constraints on other Figures of Merit such as the sheet base resistance R_{sBi}.

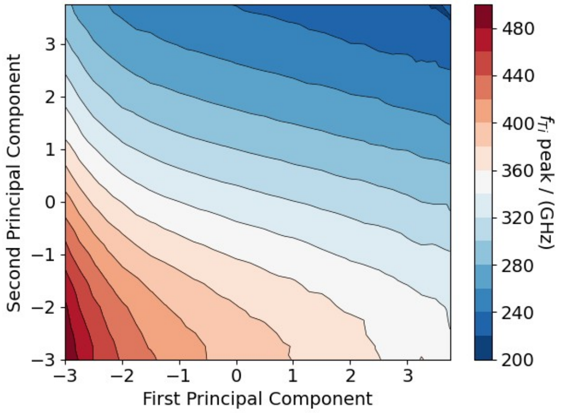

Fig.4: PCA of profile parameters with average f_{Ti}

Conclusion

Once installed, these techniques enable ambitious design goals to be achieved, freeing from silicon fabrication cycles and TCAD simulations, and using machine learning to make bipolar transistor design more sustainable, especially in terms of energy consumption.

References

[1] R. Ferguson and D. J. Roulston, "Artificial neural networks for reverse engineering bipolar transistors," *1997 21st International Conference on Microelectronics*. Proceedings, 1997, pp. 459-462.

[2] G. Caron and A. Juditsky and N. Guitard and D. Céli, "Recovery of Intrinsic Heterojunction Bipolar Transistors Profiles by Neural Networks, " *2022 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS)*. Proceedings, 2022.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory, " *Neural Comput,* 1997; 9 (8): 1735–1780.

[4] I. Sutskever and O. Vinyals and Q. Le, "Sequence to Sequence Learning with Neural Networks," *Advances in Neural Information Processing Systems*, 2014; 4.