Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.
Digital Object Identifier XX.XXXX/YYY.20XX.ZZZZZZZ

Real-Time Nonlinear Behavioral Electro-Thermal Device-Level Emulation of IGBT on Heterogeneous Adaptive Compute Acceleration Platform

BINGRONG SHANG, (Student Member, IEEE), TIANSHI CHENG, (Student Member, IEEE), TIAN LIANG, (Member, IEEE), NING LIN, (Member, IEEE), AND VENKATA DINAVAHI, (Fellow, IEEE)
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2V4, Canada

Corresponding author: Bingrong Shang (e-mail: bingrong@ualberta.ca).

This work was supported by the Natural Science and Engineering Research Council of Canada (NSERC).

ABSTRACT Power converter design evaluation by means of real-time simulation techniques is prevalent, although it is mostly restricted to simple power semiconductor switch models that exclude device-level physical details. In this work, the nonlinear high-order electro-thermal model of the Insulated-gate bipolar transistor (IGBT) is developed and then deployed onto the heterogeneous digital hardware for real-time implementation. As the complexity of the nonlinear behavioral model (NBM) of the IGBT poses a significant computational burden on real-time hardware emulation, machine learning (ML) methodology is utilized so that the trained model can reproduce the characteristics of its original counterpart as much as possible and then it is implemented on the Adaptive Compute Acceleration Platform (ACAP), which composes of the processing system (PS), programmable logic (PL), and Artificial Intelligent Engine (AIE). The vector multiplication feature of the AIE caters to mathematical operations of the ML-based model particularly well and consequently enables it to be executed in real-time with remarkable speedup over the original model with which matrix inversion is otherwise mandatory. Finally, the validation for real-time device-level results and system-level results of a multi-converter system is provided by SaberRD® and MATLAB/Simulink®.

INDEX TERMS Adaptive Compute Acceleration Platform (ACAP), AI Engine (AIE), Artificial neural network (ANN), Field-Programmable Gate Array (FPGA), Insulated-Gate Bipolar Transistor (IGBT), machine learning (ML), power electronic converters, real-time systems.

LIST OF ABBREVIATIONS

ACAP	Adaptive Compute Acceleration Platform
AIE	AI Engine
ANN	Artificial Neural Network
APU	Application Processing Unit
BRAM	Block RAM
FPGA	Field-Programmable Gate Arrays
IGBT	Insulated Gate Bipolar Transistor
MAE	Mean Absolute Error
ML	Machine Learning
NBM	Nonlinear Behavioral Model
NN	Neural Network
NoC	Network on Chip
PL	Programmable Logic
PS	Processing System
ReLU	Rectified Linear Unit
SIMD	Single Instruction Multiple Data

I. INTRODUCTION

Power electronic converters have been playing a significant role in power supply systems in many domains, such as rail transportation [1], electric vehicles [2], and ship power systems [3]. The Insulated-gate bipolar transistor (IGBT) is now one of the most important and extensively used power semiconductor switches in the aforementioned applications for its advantages and characteristics, such as large capacity, simple driving, easy protection, and high switching frequency.

There is a growing volume of literature that establishes the system-level simulation of these converter-based systems for their design and performance evaluations [4]–[6], where most of them are based on detailed modeling or average value modeling, which suffices for the testing and verification of system-level converter functions such as frequency regulation and voltage adjustment. When an in-depth study is required for a comprehensive electro-thermal transient analysis, the device-level modeling is compulsory [7], as it reveals the transient performance of the power semiconductor switch, so that the transient voltage, current, and thermal stresses can be monitored accurately for real converter design evaluation [8].

Various device-level IGBT models have been developed and widely used in the past for power converter simulation [9], [10], such as the analytical model, and the nonlinear behavioral model (NBM). However, the modeling complexity due to the inclusion of device transients poses a significant challenge accompanied by a high chance of numerical divergence. This often results in a short simulation duration that is even insufficient for the system to reach its steady state, especially in commercial simulation tools such as PSpice®, Multisim™, and SaberRD®. Therefore, hardware acceleration using FPGA has been adopted for medium-scale power converters where a dramatic speedup over CPU was attained [11], [12]. In addition, [13] implements the device-level simulation of the IGBT model using the parallel algorithm on GPU, which also significantly improves the simulation efficiency. Real-time simulation [14] is playing an increasingly vital role in the development and testing stages of power electronics and requires the model to be updated strictly within the corresponding simulation time-step, but the nonlinear property of the device model determines that real-time execution can hardly be met due to a Newton-based iterative solution of a high-order matrix equation. As a result, both hardware acceleration and algorithm optimization are necessary to achieve that goal.

Machine learning (ML) has begun to be employed in power systems and power converters to reduce the computational burden of conventional models [15], [16], and various neural networks (NNs) including gate recurrent unit (GRU) [17] and recurrent neural networks (RNN) [18] are utilized to train models and obtain accurate results and improve the simulation efficiency. As a novel and time-saving approach, ML can also be applied to the study of circuit transients by learning a specific dataset and configuring the NN to create the design-compliant models [19]. However, this approach has yet to be explored for power electronics device simulations. In this paper, the ML methodology is adopted for avoiding high-dimensional matrix equations that are challenging to solve by traditional methods.

Compared to the conventional FPGA, the Versal™ Adaptive Compute Acceleration Platform (ACAP) from Xilinx® has an innovative design in terms of hardware architecture, which combines Adaptable Engines, Scalar Engines, Intelligent Engines, and Network on Chip (NoC) to provide powerful heterogeneous acceleration for a wide range of applications [20]. As the most critical and innovative part of ACAP, the AI Engine (AIE) is a highly optimized processor with many features, such as the Single Instruction Multiple Data (SIMD) vector unit, and Very Long Instruction Word (VLIW) function that can be used in the field of real-time emulation to solve the data-intensive computing issues.

In this work, the IGBT electro-thermal NBM has been implemented and evaluated on the Versal™ ACAP's processing system (PS), programmable logic (PL), and AIE, separately. The ML-based model is proposed to accommodate the SIMD vector processing feature of the ACAP, specifically, the adoption of the NN enables faster matrix calculations to replace the complex iterative matrix inversion in the transient simulation process. The ML model is realized through learning from the dataset of IGBT NBM, and the AIE SIMD vector unit provides intrinsics [21] to make the model emulation more efficient before being implemented on the ACAP. Finally, the simulation results of a multi-converter system are verified by MATLAB/Simulink®.

This paper is organized as follows: Section II introduces the IGBT device-level nonlinear behavioral electro-thermal model. In Section III, the Versal™ ACAP architecture including PS, PL, and AIE is introduced, and the implementation and performances of the NBM in these three domains are also presented. The machine learning model, training methodology, and vectorized implementation are described in Section IV. Section V shows the validation of the ML model and hardware simulation results, and Section VI provides the conclusion.

II. NONLINEAR BEHAVIORAL ELECTRO-THERMAL DEVICE-LEVEL MODELING OF IGBT

A. IGBT NONLINEAR BEHAVIORAL MODEL

The nonlinear behavioral model [22] of an IGBT with its inherent anti-parallel diode is shown in Fig. 1 (a). According to definition,

$$i(t) = C\frac{dv(t)}{dt}, \tag{1}$$

a capacitor can be discretized by Backward Euler as:

$$\int_{t-\Delta t}^{t} i(t)\,dt = C[v(t) - v(t-\Delta t)], \tag{2}$$

$$\begin{aligned} i(t) &= \frac{C}{\Delta t}v(t) - \frac{C}{\Delta t}v(t-\Delta t) \\ &= \frac{C}{\Delta t}v(t) + I_{c_{eq}}, \end{aligned} \tag{3}$$

where Δt is the time-step. The equivalent conductance is defined as $G_{C_{eq}} = \frac{C}{\Delta t}$, and the equivalent current source $I_{c_{eq}}$=-$\frac{C}{\Delta t}v(t-\Delta t)$. Consequently, for capacitor C_{ge}, the conductance $G_{C_{ge}}$ and current source $i_{C_{geeq}}$ are given as:

$$G_{C_{ge}} = \frac{C_{ge}}{\Delta t}, \tag{4}$$

$$i_{C_{geeq}} = -G_{C_{ge}} \cdot v_{C_{ge}}(t-\Delta t). \tag{5}$$

The discretized forms of nonlinear capacitors C_{cg} and C_{ce} are identical, for example:

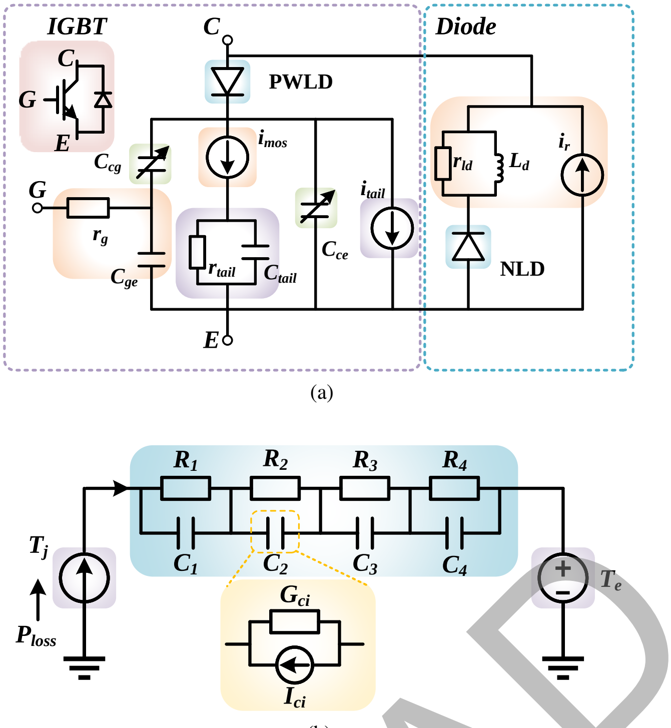
$$C_{cg} = \begin{cases} (C_{cgo} \cdot (1 + \frac{v_{C_{cg}}}{v_{C_{go}}})^{-m}), v_{C_{cg}} > 0 \\ C_{cgo}, \qquad\qquad\qquad v_{C_{cg}} \leq 0. \end{cases} \tag{6}$$

where m is the Miller capacitance exponent coefficient, which is set to 0.5 by default, and C_{cgo} is the fixed capacitance, given in Appendix A.

Similar to C_{ge}, the conductance could be calculated as $G_{C_{cg}} = \frac{C_{cg}}{\Delta t}$, and the equivalent current source as:

$$i_{C_{cgeq}} = \frac{q_{C_{cg}}(t) - q_{C_{cg}}(t - \Delta t)}{\Delta t} - G_{C_{cg}} \cdot v_{C_{cg}}(t), \tag{7}$$

where $q_{C_{cg}}$ is the charge.



(a)

(b)

FIGURE 1: (a) High-order IGBT nonlinear equivalent circuit; (b) equivalent thermal network.

Since the IGBT has three operating states: OFF state, linear, and saturation regions, the metal-oxide-semiconductor field-effect transistor (MOSFET) is adopted for model description, and its equivalent current i_{mos} can be formulated by three segments, namely

$$i_{mos} = \begin{cases} 0, (v_{C_{ge}} < V_{th}) \ \& \ (v_d \leq 0) \\ a_2 \cdot v_d^{(z+1)} - b_2 \cdot v_d^{(z+2)}, v_d < (y \cdot \Delta v_{C_{ge}})^{\frac{1}{x}} \\ \frac{\Delta v_{C_{ge}}^2}{(a_1 + b_1 \Delta v_{C_{ge}})}, others, \end{cases} \tag{8}$$

where a_1, a_2, b_1, b_2, x, y and z are coefficients, $v_{C_{ge}}$ and v_d are the voltages over capacitor C_{ge} and i_{mos}, respectively, V_{th} is the IGBT channel threshold voltage, and $\Delta V_{C_{ge}}$ is defined as

$$\Delta v_{C_{ge}} = v_{C_{ge}} - V_{th}. \tag{9}$$

consequently, the conductance G_{mosvd} and transconductance $G_{mosvcge}$ resulting from the discretization of the component can be derived by taking partial derivatives of v_d and $v_{C_{ge}}$, respectively, and each operation state has a different form.

1) ON state

Under ON state, i.e. v_d is less than the value of $(y \cdot \Delta v_{C_{ge}})^{\frac{1}{x}}$, the conductance and transconductance are expressed by the following equations

$$G_{mosvd} = \frac{\partial i_{mos}}{\partial v_d} = a_2(z+1) \cdot v_d^z - b_2(z+2) \cdot v_d^{(z+1)}, \tag{10}$$

$$G_{mosvcge} = \frac{\partial i_{mos}}{\partial v_{C_{ge}}} = \frac{\partial a_2}{\partial v_{C_{ge}}} \cdot v_d^{(z+1)} - \frac{\partial b_2}{\partial v_{C_{ge}}} \cdot v_d^{(z+2)}. \tag{11}$$

2) Transient state

Under the transient stage, the conductance G_{mosvd} is zero, and the transconductance can be derived as

$$G_{mosvcge} = \frac{2\Delta v_{C_{ge}}}{(a_1 + b_1 \Delta v_{C_{ge}})} - \frac{b_1 \Delta v_{C_{ge}}^2}{(a_1 + b_1 \Delta v_{C_{ge}})^2}. \tag{12}$$

3) OFF state

When the IGBT is OFF, both G_{mosvd} and $G_{mosvcge}$ are zero.

Taking the different forms of G_{mosvd} into consideration, the companion current of i_{mos} can be calculated by

$$I_{moseq} = i_{mos} - G_{mosvd} \cdot v_d - G_{mosvcge} \cdot V_{C_{ge}}. \tag{13}$$

The tail current I_{tail} occurs when the IGBT is being turned off, and it can be estimated using the formula below

$$I_{tail} = \begin{cases} 0, \frac{V_{tail}}{R_{tail}} < i_{mos} \\ (\frac{V_{tail}}{R_{tail}} - i_{mos}) \cdot i_{rat}, others, \end{cases} \tag{14}$$

where i_{rat} is a fixed current.

Finally, all subunits are combined and expressed as

$$\mathbf{G_{IGBT}} \cdot \mathbf{v_{IGBT}} = \mathbf{I_{IGBTeq}}, \tag{15}$$

where $\mathbf{G_{IGBT}}$ is the 5×5 admittance matrix, $\mathbf{v_{IGBT}}$ is the IGBT node voltage, and $\mathbf{I_{IGBTeq}}$ is the companion current.

B. DIODE NONLINEAR BEHAVIORAL MODEL

The nonlinear behavioral power diode model is demonstrated in the right part of Fig. 1 (a). The relationship between diode static current I_d and its junction voltage is expressed by

$$I_d = I_s \cdot [e^{(\frac{V_j}{V_b})} - 1], \tag{16}$$

where I_s is the leakage current, V_b is the junction barrier potential, and V_j is the static junction voltage.

The nonlinear diode (NLD) conductance G_j and the companion current I_{jeq} are

$$G_j = \frac{\partial I_d}{\partial V_j} = \frac{I_s}{V_b} e^{\frac{V_j}{V_b}}, \tag{17}$$

$$I_{jeq} = I_d - G_j \cdot V_j. \tag{18}$$

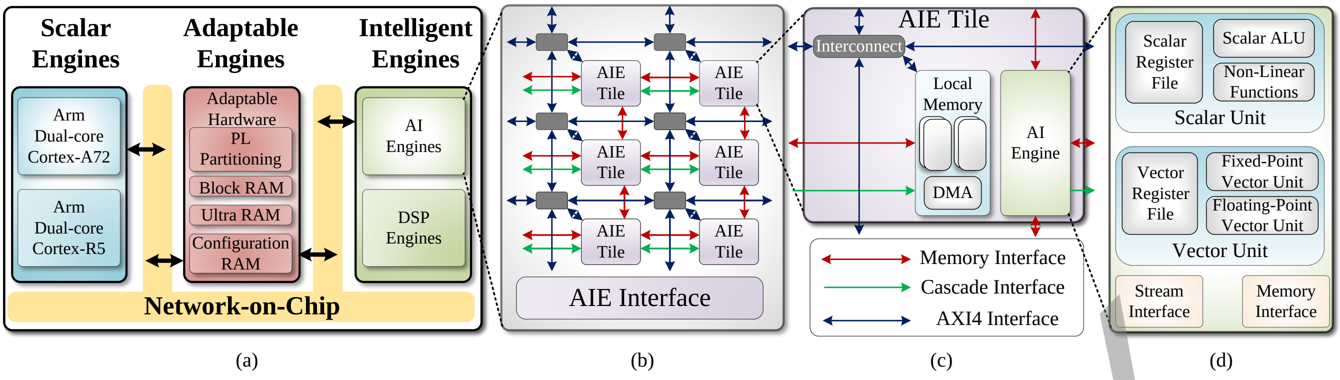


FIGURE 2: (a) Architecture of ACAP; (b) AI Engine array; (c) AI Engine tile; (d) AI Engine architecture.

C. IGBT ELECTRO-THERMAL MODEL

As given in Fig. 1 (b), the process in which the power loss causes semiconductor junction temperature rise can be modeled by the R-C pairs as an equivalent electro-thermal network [23] which is generally expressed as

$$Z_{th} = \sum_{i=1}^{N} R_{th(i)}(1 - e^{-\frac{t}{\tau_i}}), \tag{19}$$

$$C_{th(i)} = \frac{\tau_i}{R_{th(i)}}, \tag{20}$$

where $R_{th(i)}$ and τ_i are constants. The power loss of the IGBT P_{loss} is numerically equal to the input current of the transient thermal impedance equivalent circuit. On the other hand, the terminal voltage of the current source can be taken as the semiconductor's junction temperature T_j,

$$T_j(t) = \sum_{i=1}^{4} \frac{P_{loss}(t) + I_{ci}(t - \Delta t)}{G_{ci} + R_{th(i)}^{-1}} + T_e, \tag{21}$$

where T_e stands for the ambient temperature, G_{ci}=Δt/2$C_{th(i)}$, and I_{ci} is the capacitor history current.

III. IGBT NBM IMPLEMENTATION ON ACAP

Versal$^{\text{TM}}$ devices are the first ACAP based on the TSMC 7 nm FinFET process technology of Xilinx®. Fig. 2 (a) depicts the architecture of ACAP, which consists of a scalar engine (PS), an adaptable engine (PL), and an intelligent engine, all of which are connected together via a series of high-speed and integrated horizontal and vertical paths NoC to achieve remarkable performance and meet design timing, speed, and logic utilization requirements.

A. IGBT DESIGNS ON ACAP

1) AI Engine: As shown in Fig. 2 (b), the AIE array is the top-level hierarchy of the AIE architecture, which integrates a two-dimensional array of AIE tiles. The AIE array interface enables the AIE to communicate with the rest of the Versal$^{\text{TM}}$ device through the NoC or directly to the PL. The AIE tile architecture is shown in Fig. 2 (c), where each tile includes one tile interconnect module which handles AXI4 input/output, a memory module, and an engine, which can access up to 4 memory modules in four directions. The AIE, shown in Fig. 2 (d), is a highly-optimized processor that supports both fixed-point and floating-point precision and is organized as an array of AIE tiles, which can contain up to 400 tiles on the VC1902 device used in this work.

The AIE programming flow is carried out in two phases with the Vitis® integrated design environment: kernel programming and graph programming. A kernel describes a specific computing process running on a single AIE tile where C/C++ code is used for programming, and a C++ framework is provided by Xilinx to create graphs from kernels that contain declarations for the graph nodes and connections. A graph will instantiate and connect the kernels using buffers and streams, and also describe the data transfer between the AIE array and the rest of the ACAP device.

Fig. 3 shows the dataflow graph and kernels of the NBM implementation, which is achieved by 5 AIE kernels (pre_cal, $diode$, $igbt\_on$, $igbt\_off$, and $igbt\_transient$), connections, and different types of buffer, where the data transfer between kernels is memory-to-memory and the transmission of data between kernels and PL is stream-to-memory or memory-to-stream. First, the node voltage of the IGBT is sent as input to the first kernel pre_cal for parameters precalculation, the second kernel $diode$ computes the parameters of the diode, and the third to fifth kernels $igbt\_on$, $igbt\_off$, and $igbt\_transient$ are designed to perform IGBT nonlinear functions in the ON state, OFF state, and transient state, respectively, and finally, the outputs make up the admittance matrix in (15).

2) PS: As shown in the scalar engine part of Fig. 2 (a), the Application Processing Unit (APU) is based on the ARM Cortex-A72 processor core to provide general-purpose computing in a standard programming environment [24], which is chosen for IGBT NBM computation since it offers higher capabilities and a high clock frequency of up to 1700MHz. The OpenCL and the Xilinx Runtime (XRT) methodology are adopted for software programming, which enables multiple kernels to be executed concurrently with initialized command queue and thus is highly efficient in

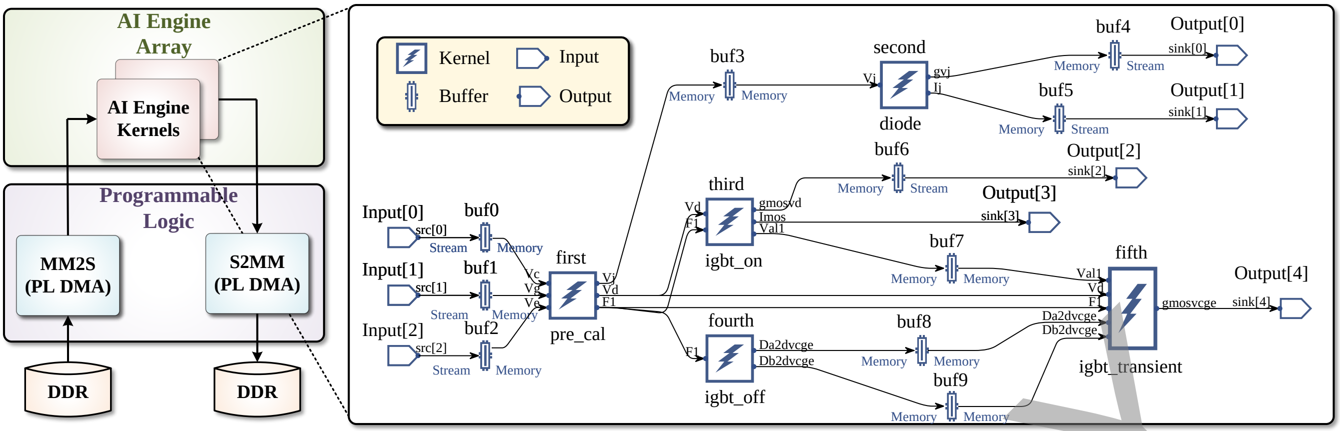


FIGURE 3: AI Engine data flow graph of IGBT NBM.

performance.

3) *PL*: PL is an extensible structure that enables the creation of a wide range of conceivable functions. It consists of DSP engines, configurable logic blocks, Configuration RAM, and Block RAM (BRAM), which can be configured together to create numerous types of hardware functionalities including accelerators, processors, functional pipeline units, and peripherals [24]. As shown in the left part of Fig. 3, PL establishes connections between PS, NoC, AIE, high-density I/O buffers, and components instantiated within the PL. In the IGBT NBM design, the GMIO port is used to connect external memory mapped to or from the global memory, which accesses DDR memory directly with a bandwidth throughput of 3200 MB/s. The connections and configuration of the PL elements are captured in the Vivado® design suite and the Vitis® unified software platform toolchain using a programmable device image.

B. COMPARISON OF NBM IMPLEMENTATIONS ON THREE DOMAINS

Fig. 4 shows the setup of the hardware platform Xilinx Versal™ VCK190 board with the ACAP device XCVC1902. The IGBT NBM is implemented on the PS, PL, and AIE of the ACAP, respectively, for a comprehensive evaluation of different design schemes. When the simulation duration is 0.05s, the actual execution time for the simulation is 0.042s on the PS. Then the real-time ratio could be expressed as $\frac{0.05s}{0.042s} = 1.19$, which indicates that for a single IGBT, the simulation speed is slightly faster than real-time. However, the simulation of a power converter with many IGBTs slows down significantly due to the inadequate scalability of PS.

Table 1 lists the latency and resource utilization of NBM implementation on AIE and PL. While the PL has the advantages of numerous resources and customizability to support the simulation of systems with multiple IGBTs, a heavy data dependency of the NBM restricts parallelism and ultimately leads to high latency. The AIE has highly optimized processors and a data stream frequency of 1GHz for efficient parallel processing. The AIE scalar processor has an excellent performance on fixed-point data processing but is not ideal for floating-point data required by NBM, as shown in Table 1. To accelerate the computing process, the ML strategy and AIE Vector Unit are adopted, as the adapted vectorized data type and SIMD features enable the IGBT NN model to be processed simultaneously.

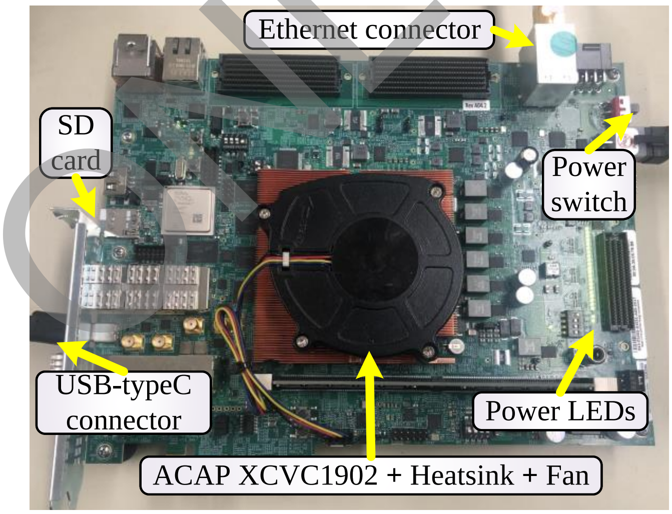


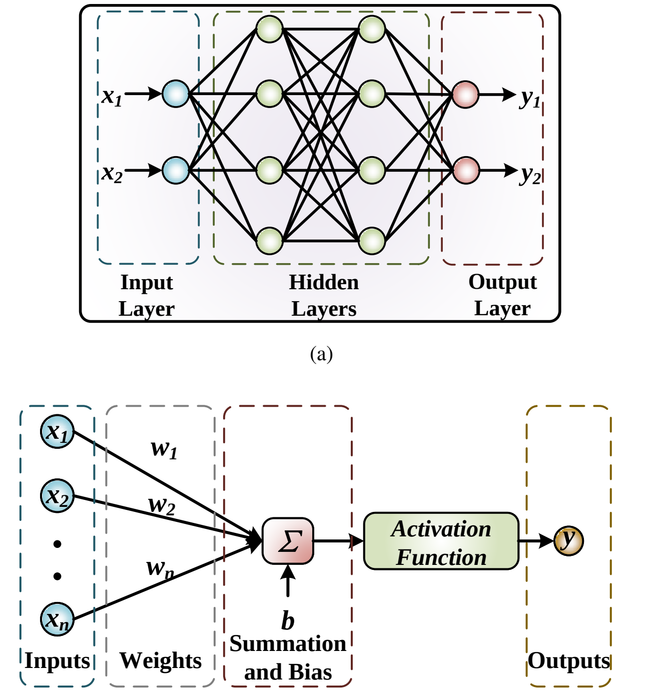
FIGURE 4: Xilinx® VCK190 board setup.

TABLE 1: NBM implementation in AIE and PL

Part	Latency	Resource Utilization		
AIE Scalar Unit	10.946μs	AIE Tiles	5	1.25%
		Kernels	5	-
PL	3.37μs	BRAM	28	1.45%
		URAM	0	0.00%
		DSP	252	12.80%
		LUT	52230	5.80%
		FF	21306	1.18%

IV. ML MODELING AND REALIZATION OF NBM

Based on the NBM performance evaluation in the previous section, it can be seen that the real-time performance is less



(a)

(b)

FIGURE 5: Neural network structure: (a) ANN basic structure with three layers; (b) neural network internal model.

than satisfactory. A machine learning-based co-simulation technique is proposed to streamline the computational procedure while maintaining simulation accuracy.

A. SELECTION OF NEURAL NETWORK TOPOLOGY

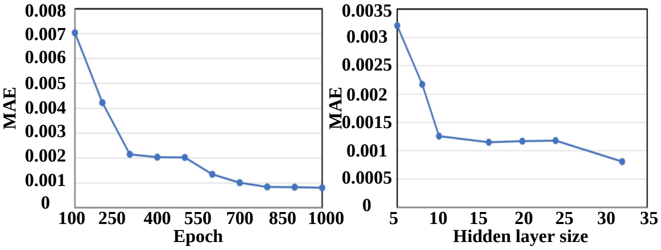
Different neural networks such as convolutional neural networks (CNN), recurrent neural networks (RNN), and artificial neural networks (ANN) are novel trends in the realm of machine learning, providing impetus for various applications. Similarly, the NN methodology can be valuable in the field of real-time simulation, as one of its benefits is that it can take advantage of the numerical prediction property to derive the corresponding output model by training on specific data, thus avoiding the extensive computations caused by iterations during transient states.

In Fig. 5 (a), an elementary version of the neural network is depicted, with a multilayer structure formed by certain neurons, notably the input layer, the hidden layer, and the output layer, each node in the upper layer is linked to all the nodes in the next layer. The mathematical expression is

$$\mathbf{Y} = \mathbf{f}(\mathbf{X} \cdot \mathbf{W} + \mathbf{b}) = \mathbf{f}(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i W_i} + \mathbf{b}), \tag{22}$$

where $\mathbf{X}$ is the input, $\mathbf{n}$ is the number of neurons, $\mathbf{Y}$ is the output, $\mathbf{W}$ is the weight, and $\mathbf{b}$ is the bias.

Fig. 5 (b) represents the general mathematical model of NN, where the input variables from x to x_i are multiplied with the weight matrix $\mathbf{W}$ and summed with the bias value $\mathbf{b}$. Finally, the activation function serves as a nonlinear mapping, limiting the amplitude of the output to a specific range. Common activation functions include Sigmoid, Tanh, and rectified linear unit (ReLU) [25], of which ReLU is the most popular type in machine learning compared to the Sigmoid and Tanh functions since ReLU has only a linear relationship and its computation is faster than the other, which needs to perform exponential operations.



(a)

(b)

FIGURE 6: IGBT ANN model's error reduction process.

In this work, ANN is chosen as the IGBT NBM transient state machine learning model because it has the feature of fitting the intermediate data curve by the first and last data only, which avoids the problem of computational iterations in traditional electromagnetic transient (EMT) models, and its high parallelism and low execution delay can match the criteria of transient simulation.

B. DATA COLLECTION AND TRAINING METHODOLOGY

One crucial part of ML training of devices is the selection of the dataset since it will influence the accuracy of the training results and the generality of the model. For the IGBT Siemens BSM300GA160D, rated 1600V, 300A in this work, where the parameters are provided in Appendix A, the dataset is extracted from the MATLAB simulation results of the IGBT NBM, and both the turn-on and turn-off data during the transient state should be of concern.

The corresponding IGBT NBM ANN model has 5 input variables including the initial and last status of the transient state voltage V_{start}, V_{end}, current I_{start}, I_{end}, and gate signal V_g. All these data are normalized to (-1,1) using min-max normalization, which allows for easier data processing and better training performance.

The mean absolute error (MAE) is used to measure the accuracy of the training model:

$$MAE = \sum_{i=1}^{n} \frac{|y_i^{pre} - y_i|}{n}, \tag{23}$$

where n is the total number of the output, y_i is i^{th} originate value from the dataset, and the y_i^{pre} is the corresponding output of the ANN model. The Adam optimization algorithm is adopted as the training methodology in this work to minimize the error [26]. Fig. 6 shows the MAE of the IGBT ANN model, which presents the error reduction during the training process. The training epoch is selected as 1000 to

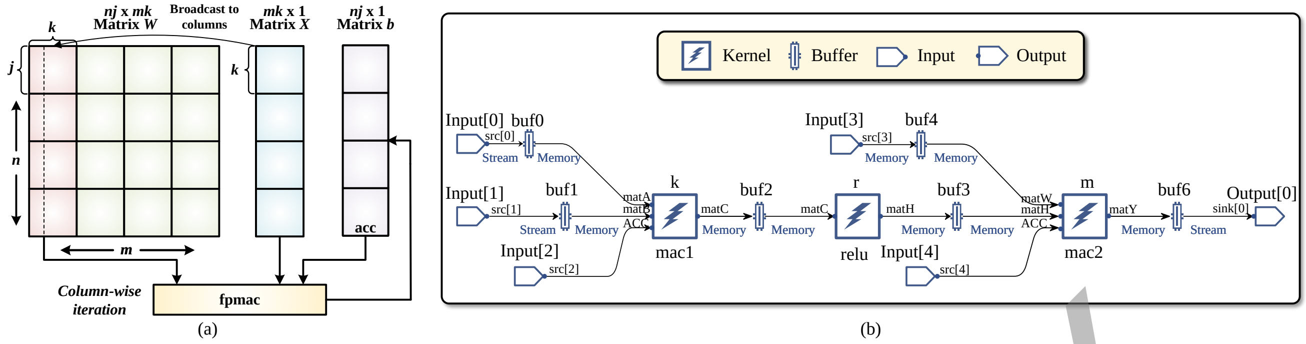


FIGURE 7: (a) Vectorized matrix multiplication in column; (b) the IGBT ANN model AIE implementation.

reduce error, and the hidden layer size is set to 32 to improve the efficiency of the AIE vector code since the size of the accumulator is a multiple of 8-bit. Since the MAE of one hidden layer is not significantly distinct from that of two hidden layers, it is used to achieve optimal performance.

C. MATRIX MULTIPLICATION IMPLEMENTATION WITH AIE

From the previous part of this section and the mathematical expression, the input variables need to be multiplied by the weight and summed by bias, which could be seen as the matrix multiplication and addition for the hidden layer and output layer. Some changes are performed to the matrix size that has no impact on the outcome to make the operations adaptable for the AIE vectorized code, for example, for the hidden layer, the size of the weight matrix $\mathbf{W}$ is 32×8, the input matrix $\mathbf{X}$ is 8×1, and the bias matrix $\mathbf{b}$ is 32×1.

The column-based matrix multiplication is implemented using vectorized AIE code, where the vector data types pack multiple scalar data elements into a wider vector. In this case, both the AIE API and intrinsics are employed to increase design productivity. The AIE API, which is implemented as a C++ header-only library and offers types and operations that are converted into effective low-level intrinsics, is a portable programming interface for accelerators. In the meantime, the vector data types and the MAC intrinsics [21] are deployed for application-level programming in this work.

There are two solutions based on AIE floating-point intrinsics to implement the matrix multiplication; the first strategy is to perform the multiplication with $fpmul$ and then add it with the bias matrix to the accumulator using $fpmac$. Another methodology, the more efficient way presented in this paper, is to apply $fpmac$ intrinsic only as shown in Fig. 7 (a). Firstly, the bias matrix $\mathbf{b}$ is loaded to the accumulator, then the weight matrix $\mathbf{W}$ is stored at several accumulators by column, and each column in the weight matrix is multiplied by the corresponding row of the input matrix $\mathbf{X}$, where the $fpmac$ intrinsic is applied to perform both the matrix multiplication and addition, the full IGBT ANN AIE vectorized matrix calculation is shown in Fig. 7 (b).

V. EMULATION RESULTS AND DISCUSSION

A. IGBT ANN MODEL VALIDATION AND PERFORMANCE

Fig. 8 gives the ANN model training results compared with the offline device-level (100 ns time-step) simulation tool SaberRD®, where Fig. 8 (a) is the IGBT transient current and voltage of the turn-on state and Fig. 8 (b) is the turn-off state. Fig. 8 (c) and (d) show the IGBT junction temperature at 200 A and 333 A, where the latter needs an additional cooling system. Table 2 shows the latency and resource consumption of different parts of the ANN model implemented in AIE. A comparison of matrix multiplication implementations on different hardware platforms is given in Table 3, for the same size matrix multiplication, AIE is 2.6 times faster than CPU and more than 28 times faster than FPGA.

TABLE 2: IGBT ANN model performance in AIE

Part	Latency	Size	Resource
Hidden layer	136 ns	$[32 \times 8] \times [8 \times 1] + [32 \times 1]$	0.5%
Output layer	1706 ns	$[80 \times 32] \times [32 \times 1] + [80 \times 1]$	0.5%
ReLU	68 ns	$[32 \times 1]$	0.25%

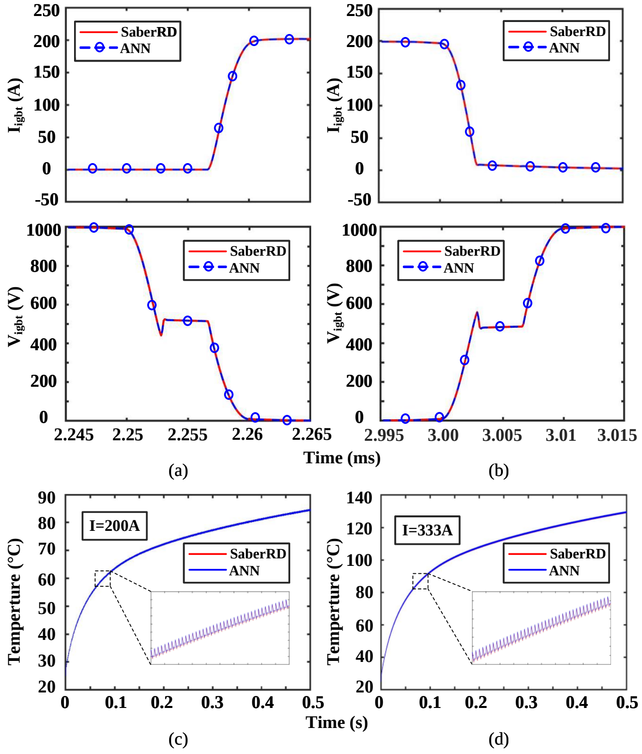
TABLE 3: Comparison of matrix multiplications on different hardware

Hardware Type	Platform	Size	Latency
AI Engine	Versal$^{\text{TM}}$ VCK190	$[32 \times 8] \times [8 \times 1] + [32 \times 1]$	136 ns
FPGA	Zynq® ZCU106	$[32 \times 8] \times [8 \times 1] + [32 \times 1]$	3860 ns
CPU	Intel® Core$^{\text{TM}}$ i7	$[32 \times 8] \times [8 \times 1] + [32 \times 1]$	360 ns

B. REAL-TIME SYSTEM-LEVEL EMULATION RESULTS

The case study system is presented in Fig. 9, where Fig. 9 (a) shows the 2-level VSC converter. For the DC side, as shown in Fig. 9 (b), there are 4 kinds of load circuits, namely half-bridge load, buck load, boost load, and full-bridge load, and Fig. 9 (c) presents the control diagram. The system parameters are given in Appendix B. The emulation of the system is implemented on the Xilinx Versal$^{\text{TM}}$ ACAP XCVC1902, where the time-step is 5 μs. Table 4 provides the hardware resources consumption and the latency of the different parts of the system.

Fig. 10 demonstrates the simulation results of the case study system with the AC side fault F at 0.4 s as shown



(a)

(b)

(c)

(d)

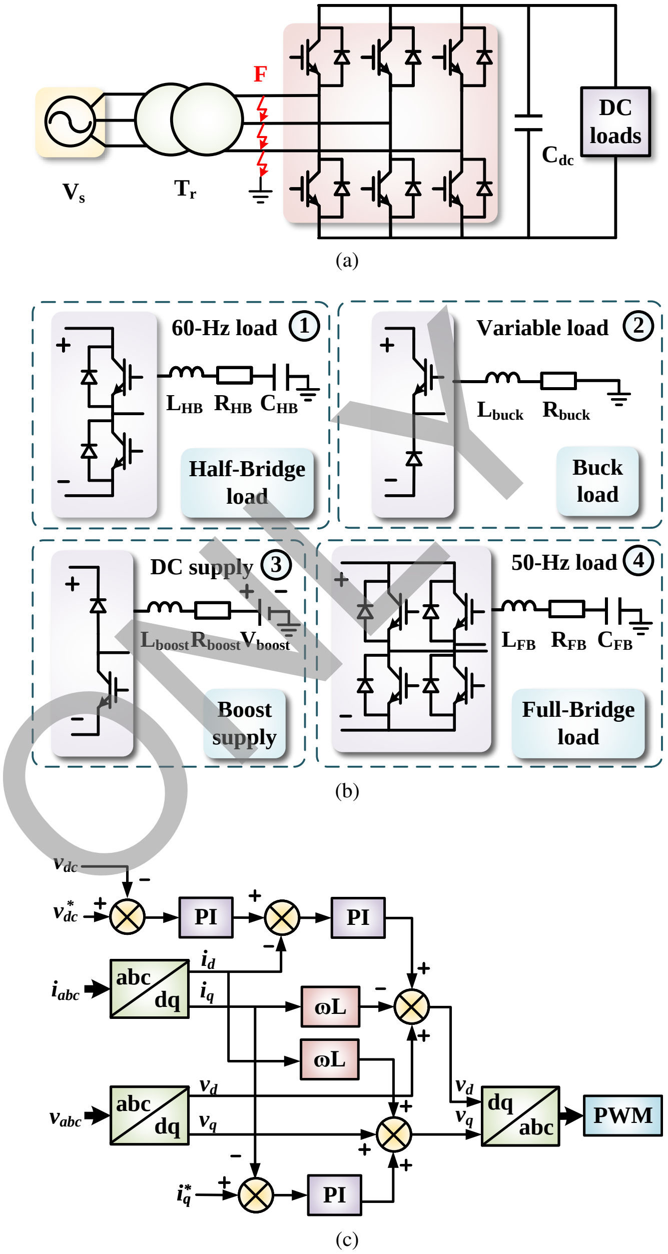
FIGURE 8: IGBT ANN model: (a)-(b) IGBT turn-on and turn-off state; (c)-(d) device junction temperature.

TABLE 4: Resources consumption of a VSC converter

Part	Latency	BRAM	DSP	FF	LUT	URAM
Control	4280 ns	0.21%	0.51%	0.20%	0.40%	0
Solver	8900 ns	0.10%	0.20%	0.28%	0.62%	0
Converter	1510 ns	0.41%	0.46%	0.24%	0.51%	0

in Fig. 9 (a). In Fig. 10 (a), before the AC side fault, the power of the grid varied in the range of approximately 600 kW to 900 kW; and it quickly drops to about 50 kW when the fault occurs. Then after 0.1 s, the grid power is gradually restored. Fig. 10 (b) displays the power of the full-bridge and half-bridge load, which both decrease from their original power at fault, and increase to peak at 0.5 s, then reinstate at 0.6 s. Fig. 10 (c) is the power of the buck load and has the same trend as the previous figures while the value drops to 0 when the fault happens. Fig. 10 (d) is the boost load power and the power remains steady before the fault, and the value changes from about -124 kW to -110 kW between 0.4 s to 0.5 s, and recovery to the original value after 0.1 s. Fig. 10 (e) and (f) is the voltage on the DC side and AC side. Fig. 11 gives the junction temperature of an IGBT in the simulation of the whole system. In Fig. 11 (a), with Cooling System 1 which has the insufficient capacity as given in Appendix A, the junction temperature reaches about 220° at the steady state. Fig. 11 (b) shows that with a decent capacity, such as Cooling System 2, the temperature remains below 70° even though the fault occurred.

In Fig. 12, the simulation results of the system are presented with the DC side half-bridge load circuit fault at 0.5 s and last for 2 seconds. Fig. 12 (a) shows the gird power between 0 and 3.0 s, and it can be seen that the power increases to about 95 kW at 0.5 s, and then returns to its original value at 2.5 s. Fig. 10 (b) is the power of the full-bridge and buck load, both of which do not change considerably after the fault occurs. In Fig. 12 (c), the power of the half-bridge load increases from its original value to 440 kW and becomes stable in the range of 390 kW to 420 kW, then restored after the fault ends at 2.5 s. Fig. 12 (d) shows the DC side voltage, which originally varied between approximately 950 V and 1040 V, and changed to between

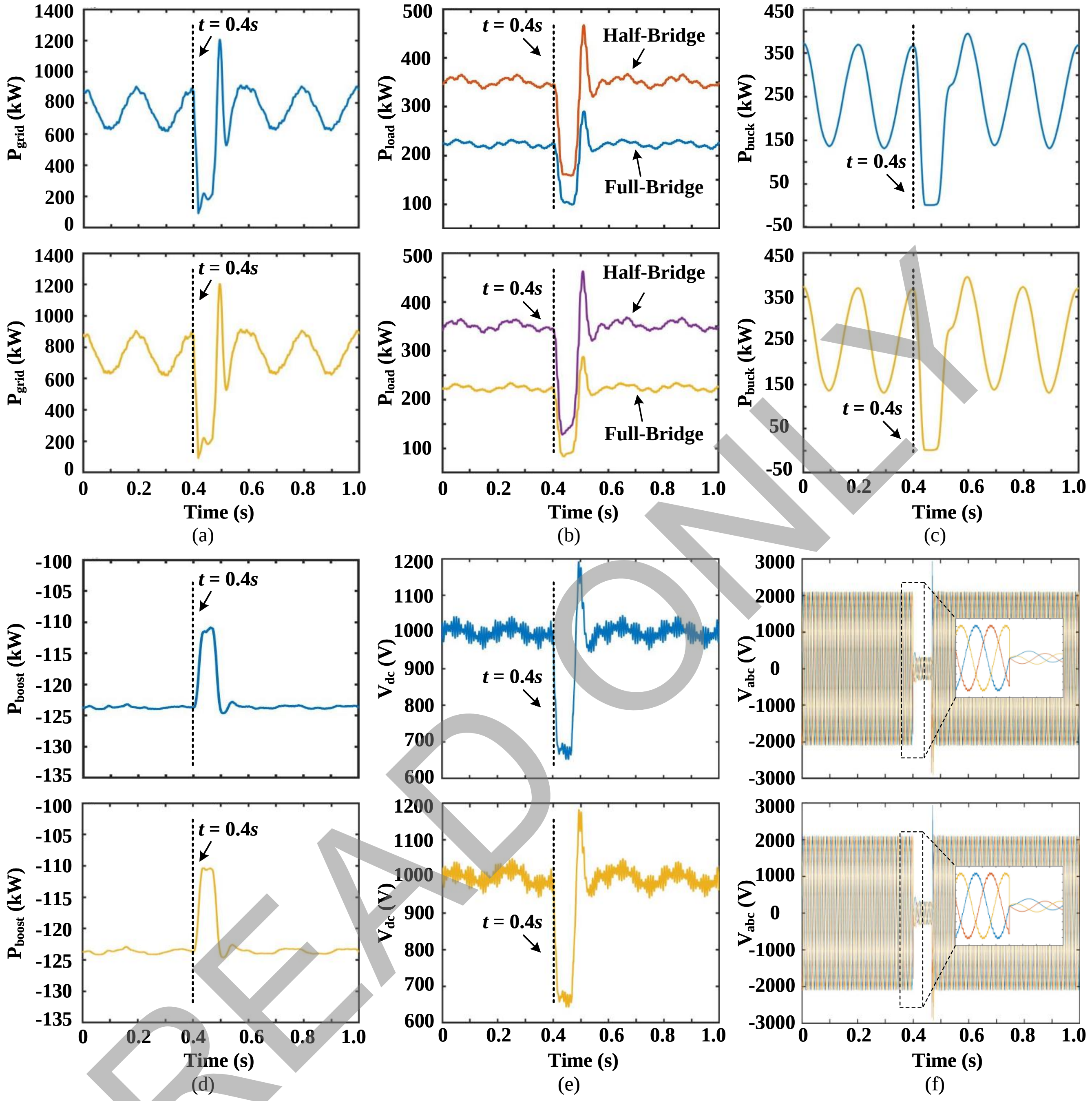


(a)

(b)

(c)

FIGURE 9: Case study of the full system: (a) AC rectifier part; (b) DC loads; (c) control diagram of 2-level VSC.



(a) (b) (c)

(d) (e) (f)

FIGURE 10: System-level results with AC fault from offline simulation (top), and ML model (bottom): (a)-(d) power of the grid, full-bridge and half-bridge loads, buck load, and boost load; (e)-(f) voltage of DC side and AC side.

940 V and 1050 V after the fault occurred.

VI. CONCLUSION

Real-time emulation of a device-level nonlinear behavioral model of IGBT is a challenging task due to its high computation burden arising from the need for an iterative solution of device equations to obtain a convergent solution of every nanosecond scale time-step. In this paper, a machine learning strategy is proposed to tackle the IGBT nonlinear behavioral electro-thermal model and demonstrated in a multi-converter supply-load system case study. The model is implemented on three main domains of a novel heterogeneous ACAP hardware: PS, PL, and AIE, which are introduced in detail in terms of functionality and features. The performance evaluation results, covering latency and hardware resource consumption, are provided separately. To make better utilization of the VCK190 hardware platform and AIE characteristics to achieve the requirements of real-time simulation, the

This article has been accepted for publication in IEEE Open Journal of the Industrial Electronics Society. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJIES.2022.3220248

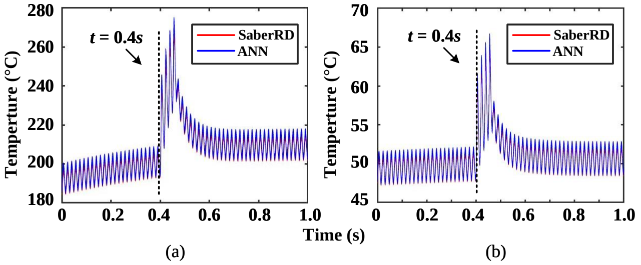


FIGURE 11: Device junction temperature with: (a) Cooling System 1; (b) Cooling System 2.

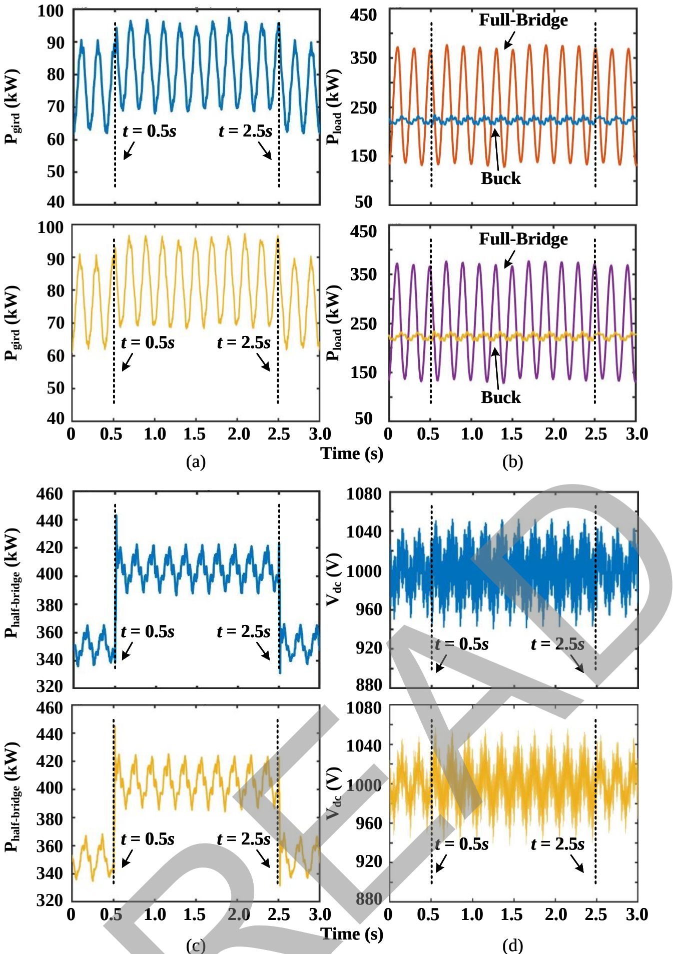


FIGURE 12: System-level results with half-bridge load circuit fault from offline simulation (top), and ML model (bottom): (a)-(c) power of the grid, full-bridge and buck load, and half-bridge load; (d) DC side voltage.

IGBT ML-based model and NNs training methodology are proposed, where the ANN model is adopted to convert the complex computational iterative process of the transient state into the simpler matrix operations. From results comparisons with the conventional model in device-level emulation, the error of the IGBT ML model is within 1%, and the real-time requirement can be achieved with less resource consumption. The system-level simulation results are given for two different fault scenarios on both AC and DC sides and validated by MATLAB/Simulink®. The proposed modeling and implementation strategies can be applied in the future for real-time emulation of energy conversion systems in various practical applications.

APPENDIX A

The parameters of the IGBT Siemens BSM300GA160D, rated 1600V, 300A behavioral model:
V_t = 6.3 V, x = 0.974, y = 1.429, z = 0.369, a_1 = 0.022, b_1 = 0.004, a_2 = 92.5129, b_2 = 4.0188, r_{tail} = 1 $\mu\Omega$, C_{tail} = 10 F, i_{rat} = 0.05, C_{geo} = 40 nF, C_{cgo} = 110 nF.

Cooling System 1: R_1 = 2.1 K/kW, R_2 = 9.2 K/kW, R_3 = 42.6 K/kW, R_4 = 6.3 K/kW, τ_1 = 0.0008 s, τ_2 = 0.013 s, τ_3 = 0.05 s, τ_4 = 0.063 s.

Cooling System 2: R_1 = 1.33 K/kW, R_2 = 7.05 K/kW, R_3 = 5.23 K/kW, R_4 = 2.8 K/kW, τ_1 = 0.00147 s, τ_2 = 0.034 s, τ_3 = 0.168 s, τ_4 = 1.11 s.

APPENDIX B

The parameters of the case study system:
The grid voltage V_s = 490 V (L-L), 60 Hz; the transformer 1MVA, 25 kV/490 V; C_{dc} = 0.0333 F; the half-bridge load 400+j50 kVA; the buck load 250 kW, duty D = 0.55; the boost supply V_{boost} = 500 V, duty D = 0.8; the full-bridge load 200+j50 kVA.

REFERENCES

[1] D. Ronanki and S. S. Williamson, “Modular multilevel converters for transportation electrification: Challenges and opportunities,” IEEE Trans. Transport. Electrific., vol. 4, no. 2, pp. 399–407, Jan. 2018.
[2] Q. Sun, J. Wu, C. Gan, J. Si, J. Guo, and Y. Hu, “Cascaded multiport converter for SRM-Based hybrid electrical vehicle applications,” IEEE Trans. Power Electron., vol. 34, no. 12, pp. 11 940–11 951, Apr. 2019.
[3] A. Francés-Roger, A. Anvari-Moghaddam, E. Rodríguez-Díaz, J. C. Vasquez, J. M. Guerrero, and J. Uceda, “Dynamic assessment of cots converters-based dc integrated power systems in electric ships,” IEEE Trans. Ind. Informat., vol. 14, no. 12, pp. 5518–5529, Feb. 2018.
[4] S. Horiuchi, K. Sano, and T. Noda, “An inverter model simulating accurate harmonics with low computational burden for electromagnetic transient simulations,” IEEE Trans. Power Electron., vol. 36, no. 5, pp. 5389–5397, May. 2021.
[5] A. Hadizadeh, M. Hashemi, M. Labbaf, and M. Parniani, “A matrix-inversion technique for FPGA-Based real-time EMT simulation of power converters,” IEEE Trans. Ind. Electron., vol. 66, no. 2, pp. 1224–1234, Feb. 2019.
[6] X. Meng, J. Han, J. Pfannschmidt, L. Wang, W. Li, F. Zhang, and J. Belanger, “Combining detailed equivalent model with switching-function-based average value model for fast and accurate simulation of MMCs,” IEEE Trans. Energy Convers., vol. 35, no. 1, pp. 484–496, Mar. 2020.
[7] N. Lin and V. Dinavahi, “Detailed device-level electrothermal modeling of the proactive hybrid hvdc breaker for real-time hardware-in-the-loop simulation of dc grids,” IEEE Trans. Power Electron., vol. 33, no. 2, pp. 1118–1134, Mar. 2018.
[8] H. Bai, C. Liu, E. Breaz, K. Al-Haddad, and F. Gao, “A review on the device-level real-time simulation of power electronic converters: Motivations for improving performance,” IEEE Ind. Electron. Mag., vol. 15, no. 1, pp. 12–27, Mar. 2021.
[9] L. Han, L. Liang, Y. Kang, and Y. Qiu, “A review of SiC IGBT: Models, fabrications, characteristics, and applications,” IEEE Trans. Power Electron., vol. 36, no. 2, pp. 2080–2093, Feb. 2021.
[10] K. Sheng, B. Williams, and S. Finney, “A review of IGBT models,” IEEE Trans. Power Electron., vol. 15, no. 6, pp. 1250–1266, Nov. 2000.

[11] N. Lin, B. Shi, and V. Dinavahi, "Non-linear behavioural modelling of device-level transients for complex power electronic converter circuit hardware realisation on FPGA," IET Power Electron., vol. 11, no. 9, pp. 1566–1574, Jun. 2018.

[12] H. Bai, C. Liu, R. Ma, D. Paire, and F. Gao, "Device-level modelling and FPGA-based real-time simulation of the power electronic system in fuel cell electric vehicle," IET Power Electronics, vol. 12, no. 13, pp. 3479–3487, Nov. 2019.

[13] C. Lyu, N. Lin, and V. Dinavahi, "Device-level parallel-in-time simulation of mmc-based energy system for electric vehicles," IEEE Trans. Veh. Technol., vol. 70, no. 6, pp. 5669–5678, Jun. 2021.

[14] V. Dinavahi and N. Lin, Real-Time Electromagnetic Transient Simulation of AC-DC Networks. Wiley-IEEE Press, Jun. 2021.

[15] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, "A review of machine learning approaches to power system security and stability," IEEE Access, vol. 8, pp. 113 512–113 531, Jun. 2020.

[16] G. Rojas-Dueñas, J.-R. Riba, and M. Moreno-Eguilaz, "A deep learning-based modeling of a 270 V-to-28 V DC-DC converter used in more electric aircrafts," IEEE Trans. Power Electron., vol. 37, no. 1, pp. 509–518, Jan. 2022.

[17] F. Zhang, Q. Liu, Y. Liu, N. Tong, S. Chen, and C. Zhang, "Novel fault location method for power systems based on attention mechanism and double structure GRU neural network," IEEE Access, vol. 8, pp. 75 237–75 248, 2020.

[18] X. Fu, S. Li, and I. Jaithwa, "Implement optimal vector control for LCL-Filter-Based Grid-Connected converters by using recurrent neural networks," IEEE Trans. Ind. Electron., vol. 62, no. 7, pp. 4443–4454, Jul. 2015.

[19] S. Zhang, T. Liang, and V. Dinavahi, "Machine learning building blocks for real-time emulation of advanced transport power systems," IEEE Open J. Power Electron., vol. 1, pp. 488–498, Nov. 2020.

[20] Xilinx. Inc. Versal: The first adaptive compute acceleration platform (acap). (2020, Sep.). [Online]. Available: https://docs.xilinx.com/v/u/en-US/wp505-versal-acap

[21] Xilinx. Inc. Ai engine intrinsics. (2021). [Online]. Available: https://www.xilinx.com/htmldocs/xilinx2021_2/aiengine_intrinsics/intrinsics/index.html

[22] A. Courtay, "MAST power diode and thyristor models including automatic parameter extraction," SABER User Group Meeting. Brighton, U.K., pp. 1–10, Sep. 1995.

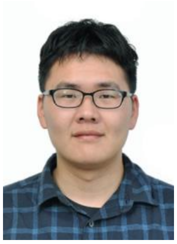
[23] R. Wu, H. Wang, K. B. Pedersen, K. Ma, P. Ghimire, F. Iannuzzo, and F. Blaabjerg, "A temperature-dependent thermal model of IGBT modules suitable for circuit-level simulations," IEEE Trans. Ind. Appl., vol. 52, no. 4, pp. 3306–3314, July-Aug 2016.

[24] Xilinx. Inc. Versal acap technical reference manual. (2022, Apr.). [Online]. Available: https://docs.xilinx.com/r/en-US/am011-versal-acap-trm

[25] J. Pomerat, A. Segev, and R. Datta, "On neural network activation functions and optimizers in relation to polynomial regression," in 2019 IEEE International Conference on Big Data (Big Data), pp. 6183–6185, Dec. 2019.

[26] R. Zaheer and H. Shaziya, "A study of the optimization algorithms in deep learning," in 2019 Third International Conference on Inventive Systems and Control (ICISC), pp. 536–539, Jan. 2019.

BINGRONG SHANG (Student Member, IEEE) received the B.Sc. degree from Henan University, Kaifeng, China, in 2020. She is currently working toward the M.Sc degree with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. Her research interests include electromagnetic transient simulation of power systems and power electronics, real-time emulation, and adaptive compute acceleration using heterogeneous hardware.



TIANSHI CHENG (Student Member, IEEE) received the B.Eng. degree in electrical engineering and automation from Southeast University, China, in 2017. From 2017 to 2018, he was a substation automation engineer of NARI Group Corporation (State Grid Electric Power Research Institute), China. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Alberta, Canada. His research interests include electromagnetic transient simulation, heterogeneous high-performance computing, real-time simulation, parallel processing, microgrid and power electronics.

TIAN LIANG (Member, IEEE) received the B.Eng. degree in electrical engineering from Nanjing Normal University, Nanjing, Jiangsu, China, in 2011, the M.Eng. degree from Tsinghua University, Beijing, China, in 2014, the Ph.D. degree in energy systems from the University of Alberta, Edmonton, AB, Canada, in 2020. His research interests include real-time simulation of power systems, power electronics, artificial intelligence, field-programmable gate arrays, and system on chip.

NING LIN (Member, IEEE) received B.Sc. and M.Sc. degrees in Electrical Engineering from Zhejiang University, Hangzhou, China, in 2008 and 2011, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Alberta, Edmonton, AB, Canada, in 2018. From 2011 to 2014, he was an engineer of substation automation, flexible AC transmission systems, and high-voltage direct current transmission control and protection. Currently, he is a senior power systems consultant. His research interests include AC/DC grids, electromagnetic transient simulation, real-time simulation, transient stability analysis, heterogeneous high-performance computing of power systems and power electronics.

VENKATA DINAVAHI (Fellow, IEEE) received the B.Eng. degree in electrical engineering from Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, in 1993, the M.Tech. degree in electrical engineering from the Indian Institute of Technology (IIT) Kanpur, India, in 1996, and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Ontario, Canada, in 2000. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada. He is a Fellow of the Engineering Institute of Canada. His research interests include real-time simulation of power systems and power electronic systems, electromagnetic transients, device-level modeling, large-scale systems, and parallel and distributed computing.