IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 22, 2021, accepted March 8, 2021, date of publication March 17, 2021, date of current version March 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3066335

Impact of NCFET on Neural Network Accelerators

GEORGIOS ZERVAKIS[1], IRAKLIS ANAGNOSTOPOULOS[2], (Member, IEEE), SAMI SALAMIN[1], (Student Member, IEEE), YOGESH S. CHAUHAN[3], (Fellow, IEEE), JÖRG HENKEL[1], (Fellow, IEEE), AND HUSSAM AMROUCH[4], (Member, IEEE)

[1]Chair for Embedded Systems, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
[2]School of Electrical, Computer and Biomedical Engineering, Southern Illinois University, Carbondale IL 62901, USA
[3]Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India
[4]Chair of Semiconductor Test and Reliability, University of Stuttgart, 70174 Stuttgart, Germany

Corresponding author: Georgios Zervakis (georgios.zervakis@kit.edu)

This work was supported in part by the German Research Foundation (DFG) through the Project "Approximate Computing aCROss the System Stack (ACROSS)."

ABSTRACT This is the first work to investigate the impact that Negative Capacitance Field-Effect Transistor (NCFET) brings on the efficiency and accuracy of future Neural Networks (NN). NCFET is at the forefront of emerging technologies, especially after it has become compatible with the existing fabrication process of CMOS. Neural Network inference accelerators are becoming ubiquitous in modern SoCs and there is an ever-increasing demand for tighter and tighter throughput constraints and lower energy consumption. To explore the benefits that NCFET brings to NN inference regarding frequency, energy, and accuracy, we investigate different configurations of the multiply-add (MADD) circuit, which is the core computational unit in any NN accelerator. We demonstrate that, compared to the baseline 7nm FinFET technology, its negative capacitance counterpart reduces the energy by 55%, without any frequency reduction. In addition, it enables leveraging higher computational precision, which results to a considerable improvement in the inference accuracy. Importantly, the achieved accuracy improvement comes also together with a significant energy reduction and without any loss in frequency.

INDEX TERMS Emerging technology, low-power, neural networks, neural processing units.

I. INTRODUCTION

Technology scaling is reaching limits in which scaling down further the operating voltage of circuits (V_{DD}) with new technology nodes is becoming profoundly difficult if not impossible. This is because V_{DD} scaling is strictly governed by the sub-threshold swing of transistors (SS), which is fundamentally limited, in conventional CMOS technology, to 60 $mV/decade$ at room temperature (300K). This is caused by "Boltzmann tyranny" [1] that dictates the distribution of charge carriers at the source of the transistor. The switching speed of any circuit is mainly determined by the ON current (I_{ON}) of the constituent transistors that form the circuit critical paths. I_{ON}, in turn, is proportional to the transistor voltage overdrive, which is V_{DD} above the threshold voltage (V_T), i.e., $I_{ON} \propto (V_{DD} - V_T)$. Therefore, in order to scale down V_{DD}, while still maintaining the same frequency, V_T must be proportionally reduced. However, the reduction in V_T increases exponentially the leakage current of transistors, resulting in a significant and unacceptable rise in the static power of circuits [2].

As a result of the discontinuation of V_{DD} scaling, improvements in computational speed and energy efficiency have become very challenging to achieve with every new generation. When it comes to speed, the frequency of circuits is not improving anymore, even though the underlying technology allows that, in order to prevent unsustainable on-chip temperatures caused by excessive power densities. Static and dynamic power, on the other hand, have exponential and quadratic dependencies on V_{DD}, respectively. Hence, the non-scalable V_{DD} will strongly restrict the potential improvements that upcoming technologies might bring to the efficiency of the circuits.

Negative Capacitance Field-Effect Transistor (NCFET) is one of the promising emerging technologies that aims to increase the steepness of the current of transistors towards pushing SS beyond its fundamental limit of 60 $mV/decade$. Among many proposed approaches in the last decade,

The associate editor coordinating the review of this manuscript and approving it for publication was Kuan Chee.

NCFET is at the forefront. This is because NCFET has recently become fully compatible with the existing CMOS fabrication process [3], which paves the way for NCFET to be adopted by the semiconductor industry for commercial usages. NCFET integrates a ferroelectric (FE) layer inside the transistor gate stack that provides an internal voltage amplification, which enables transistors to operate at a reduced V_{DD} without any loss in their switching speed. This, in turn, has a far-reaching impact on circuits' efficiency as will be later demonstrated.

Neural Network (NN) Inference: Hardware accelerators for NN inference are rapidly becoming an integral part of system-on-chips (SoCs). They typically comprise large arrays of multiply-add (MADD) circuits [4] providing a considerable increase in inference speed. In order to meet tighter and tighter throughput constraints, state of the art traditionally trades-off inference speed with accuracy in which the precision of MADD circuits is reduced (i.e., achieve higher frequency) at the cost of some accuracy loss [4], [5]. However, to satisfy such tight throughput constraints, NN accelerators integrate thousands of MADD units [4] resulting in a significant increase in energy consumption, which might not be tolerated.

Bringing NCFET and Neural Networks Together: In this work, we are the first to investigate the advantages that NCFET technology brings to NN inference w.r.t efficiency and accuracy improvements. Compared to FinFET (i.e., conventional counterpart technology), NCFET allows MADD units (*with the same precision*) to either 1) improve the speed by increasing the frequency up to 36% without any increase in the energy (i.e., effectively addressing tight throughput constraints challenge), or 2) reduce the energy up to 55% without any decrease in the frequency (i.e., effectively addressing the significant energy consumption associated with NN inference accelerators). Both improvements in either speed or energy stem from the inherent ability of negative capacitance to boost the internal transistors' voltage and hence, enable V_{DD} scaling without degrading the electrical proprieties of transistors.

Alternatively, the aforementioned improvements in NCFET can be translated into a higher precision in MADD circuits leading to a considerable increase in the inference accuracy. This comes together with lower energy consumption and, importantly, without any loss in the speed (i.e., sustaining the same frequency that the baseline FinFET provides). To evaluate the impact of NCFET-induced higher-precision MADD units on the NN inference accuracy, we study five image classification NNs [6]–[10] trained on the ImageNet dataset [11] and one audio classification NN [12] trained on the UrbanSound8K dataset [13]. In our evaluation, we studied several scenarios w.r.t different MADD precision levels showing how NCFET *always improves the accuracy and energy efficiency of NN inference*. For instance, compared to a baseline FinFET that employs 7-bit MADD circuits, NCFET increases the precision to 10-bits along with 18% energy reduction and without any loss in frequency. Such an increase in the MADD precision boosts the inference accuracy by 1.4x, on average, for the different NNs that we studied. In this work we focus on state-of-the-art Convolutional NNs (CNNs) in order to demonstrate the benefits of the NCFET. In that way, we can provide important insight regarding the next generation of NN accelerators. However, our work is not limited only to CNNs. It can also be applied to other types of NN architectures with similar results.

Our novel contributions within this article are as follows:

(1) This is the first work to provide a holistic investigation of how NCFET improves, in multiple aspects, the speed and energy efficiency of MADD circuits employed to accelerate neural network inference. We demonstrate how such improvements offered by NCFET allow the usage of MADD circuits at a higher precision along with less energy and without any loss in frequency compared to conventional FinFET (i.e., no tradeoffs). In addition, we further investigate the consequences of having higher precision MADD circuits on increasing the accuracy of various neural networks.

(2) For accurate modeling and investigations, our implementation traverses several abstracting levels, starting from physics/device level (where the negative capacitance effect does originate) all the way up to the system level (where the accuracy of neural networks is ultimately improved) through the circuit level where the efficiency (frequency and energy) as well as precision of MADD circuits are increased.

II. NEGATIVE CAPACITANCE TRANSISTORS (NCFET)

NCFET incorporates a ferroelectricity property inside the transistor gate stack, which behaves as a negative capacitance under certain conditions and hence provides an internal voltage amplification. The latter enables the *SS* of the transistor to go beyond 60 $mV/decade$ and therefore the same ON current can be achieved but at a lower V_{DD}, without the need to reduce V_T (i.e., no increase in the leakage power). In practice, NCFET does not add any additional new layer to the existing transistor but, instead, it dopes the hafnium (HfO_2)-based material – which is widely used in exiting CMOS technologies to grow high-κ dielectrics – with zirconium to create ferroelectricity [14]. Hence, NCFET does not come with any additional area overhead. Instead, NCFET comes with an overhead w.r.t the total gate's capacitance, as will be later explained.

In summary: Compared to conventional FET transistors, NCFET transistors achieve the same ON current but at a lower V_{DD} without any increase in the OFF current. Therefore, NCFET-based circuits, compared to the corresponding baseline circuits (i.e., circuits implemented using the counterpart conventional FET technology) operate at a lower V_{DD} and hence consume less power without any loss in the speed because the same operating frequency can be still sustained.

Impact of Negative Capacitance on Transistors: The presence of a FE inside the transistor gate stack can be represented using the equivalent capacitance divider that consists of the FE capacitance (C_{fe}) and the underlying internal FET capacitance (C_{FET}). The voltage amplification (A_V) at the internal

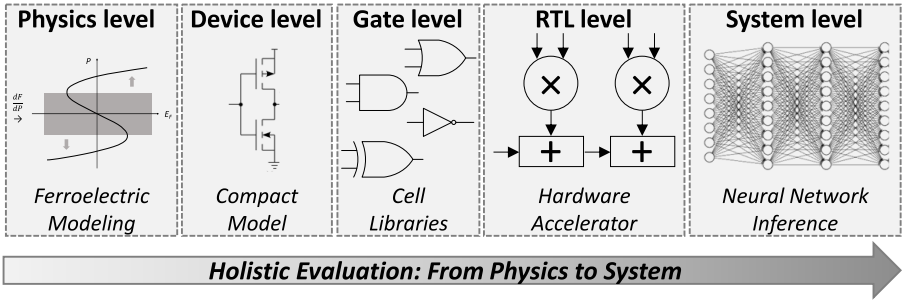

FIGURE 1. Investigating from physics all the way up to the system level the impact of NCFET on the accuracy and efficiency of neural network inference.

gate can be then expressed as shown in Eq. 1. The consequences of NC on circuits can be summarized as follows:
(1) The internal amplification will boost the flowing carriers inside the NCFET's channel and hence a higher ON current I_{ON} will be provided. Therefore, NCFET circuits will exhibit a smaller delay and thus can be clocked at higher frequency without the need to increase V_{DD}. Alternatively, the V_{DD} of NCFET circuits can be scaled down while the same frequency as in the baseline FinFET can still be sustained.
(2) Because C_{fe} has a negative value and the condition ($|C_{fe}| > C_{FET}$) must be always met to ensure no hysteresis during operation, the total capacitance of NCFET is always larger than the capacitance of the baseline FET (C_{FET}), as Eq. 3 clarifies. Therefore, compared to baseline FinFET, NCFET circuits will consume a higher dynamic power ($P_{dynamic}$) at the same V_{DD}. However, when the V_{DD} of NCFET circuit is scaled down (as explained above), a quadratic saving in $P_{dynamic}$ is obtained, which compensates to a large degree the side effect of NC on increasing the total capacitance. For example, for the NCFET circuits examined in Section V, the voltage needs to be decreased from 0.7V to 0.5V to achieve lower dynamic power than the FinFET ones at the same operating frequency.

$$A_V = \frac{\partial V_{int}}{\partial V_G} = \frac{|C_{fe}|}{|C_{fe}| - C_{FET}}; |C_{fe}| > C_{FET} \Rightarrow A_V > 1 \tag{1}$$

$$V_{int} = A_{avg} \times V_G; A_{avg} = \frac{1}{V_G}\int_0^{V_G} A_V \, dV_G \tag{2}$$

$$C_{NCFET} = \frac{C_{fe} \cdot C_{FET}}{C_{fe} + C_{FET}} > C_{FET} \Rightarrow \text{larger } P_{dynamic} \tag{3}$$

In addition to NCFET, there are several steep slope devices already presented in the literature (e.g., Tunnel FET [15], HyperFET [16], etc). Although Tunnel FET promises to offer very steep sub-threshold slope and lower leakage than the to conventional CMOS technology, it suffers from lower ON current due to the inherent nature of transport mechanism [15]. HyperFET can provide steep slope but will always have lower I_{ON} due to series resistance [16]. The main characteristic of NCFET technology, compared to other existing steep-slope transistors, is the capability of NCFET to still deliver a high ON current. Therefore, NCFET transistors can still provide a high switching speed and hence, the implemented circuits can still be clocked at a high frequency (i.e., circuit's performance is not sacrificed). This is essential for NN accelerators, as inference speed cannot be compromised [5].

III. IMPACT OF NCFET ON NEURAL NETWORK INFERENCE ACCELERATORS

Neural Network (NN) inference is one of the most common and computationally intensive workloads of today's computing systems. The core arithmetic operation performed by NNs during inference is the multiply-addition (MADD) operation. Particularly, the convolution and fully connected layers of NNs perform millions of multiplications and additions [6]. As a step to enable faster inference, significant interest is shown in the design of custom ASIC NN accelerators [4], [5], [17] targeting both cloud platforms and mobile SoCs. In addition, quantization [18] is leveraged to further improve the inference efficiency. During quantization, both weights and activations are converted to lower-precision numerical representations (e.g., perform INT8 computations in place of FLOAT32). Hence, the required precision of the MADD operations is reduced leading to significant speed improvement. The heart of NN accelerators comprises large arrays of MADD circuits [4], [5]. Considering that NNs become deeper, these accelerators need to integrate thousands or even more MADD circuits [4]. Therefore, in our analysis, we use the MADD circuit as our driving circuit to assess the figure of merit (speed, energy, and precision) of NN accelerators [19].

In this work, we perform a thorough evaluation to examine the impact of the NCFET on improving the speed, energy efficiency and accuracy of NN accelerators. As illustrated in Fig. 1, our evaluation employs a holistic approach that starts from physics (where the NC effect originates) and expands all the way up to the system level (where the NN accuracy is affected). For varying precision MADD circuits, we consider two scenarios: 1) *speed optimization*, in which we explore the frequency gain delivered by NCFET under the same energy that the baseline FinFET MADD consumes and 2) *energy optimization*, in which we explore the energy saving

achieved by the NCFET when NCFET MADD operates at the maximum frequency that the baseline FinFET MADD achieves.

IV. BRIDGING THE GAP BETWEEN PHYSICS AND SYSTEM LEVEL: OUR IMPLEMENTED SETUP

In this section, we explain the experimental setup that we implemented to evaluate the impact of the NCFET on improving the frequency and energy efficiency of MADD circuits at various precision levels and how the accuracy of NNs will be ultimately impacted. Fig. 2 demonstrates our implementation at every abstraction level to achieve our goal.

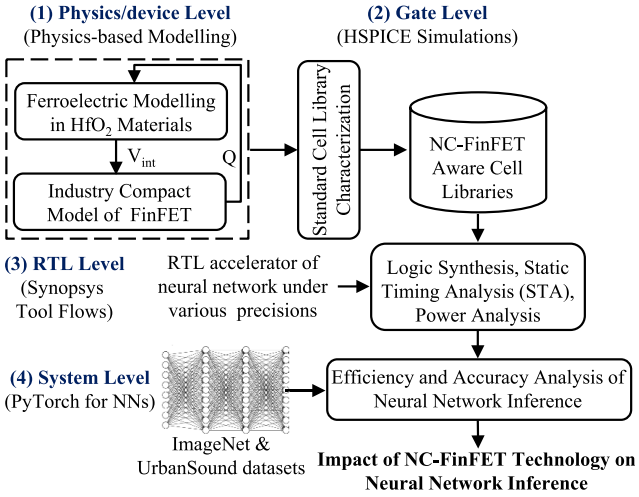

FIGURE 2. Our implementation across the computing stack starting from physics/device modeling, where the negative capacitance effect originates, all the way up to the system level, where the accuracy and efficiency of Neural Network (NN) inference are ultimately impacted.

A. NCFET MODELING: FROM PHYSICS TO CIRCUIT LEVEL

In our work, we consider NCFET with a configuration of metal-ferroelectric-metal-insulator-semiconductor (MFMIS). As our baseline technology, we employ the 7 *nm* FinFET. Hereafter, we name NC-FinFET the direct Negative Capacitance counterpart of our baseline FinFET technology. The baseline FinFET is modeled by the industry-standard compact model (BSIM-CMG) in order to accurately account for both short-channel and quantum-mechanical effects in sub-10nm geometries [20]. The FE is modeled as Eq. 4 shows based on Taylor series following the Landau-Khalatnikov (L-K) theory [21], [22]. Our modeling of polarization with the electric field accounts for the higher dielectric constant of the FE, which is crucial to correctly model how NC increases the total gate capacitance of FinFET.

$$V_{fe} = t_{fe}(2\alpha Q + 4\beta Q^3) \tag{4}$$

V_{fe} is the voltage across FE layer, Q is the gate charge, t_{fe} is the thickness of ferroelectric. α and β are FE material dependent parameters [21], [23]. The complete modeling of NC-FinFET is implemented inside the Verilog-A code of the BSIM-CMG itself and *solved in a self-consistent manner* using the commercial HSPICE simulator [24] for each bias point, ensuring charge equality and Kirchhoff's laws. For the baseline FinFET, we employ the 7 *nm* Process Design Kit (PDK) [25] where the device parameters are: $L = 21$nm, $H_{FIN} = 32$ *nm*, $T_{FIN} = 6.5$ *nm* and $EOT = 1$ *nm* [26]. For the FE, we employ Al doped HfO_2 material because it is a compatible material with the standard CMOS [14]. The α and β required in the L-K equation are calculated by the use of the remnant polarization (P_r) of 5 μC/cm^2 and the coercive field (E_c) of 1 MV/cm as extracted from the experimental measurements provided in [14]. Further details in device modeling and device parameters are available in [26].

The developed Verilog-A code of NC-FinFET is then employed within a commercial cell library characterization [24] to create NCFET-aware cell libraries. Note that the used netlists of standard cells in the characterization process are obtained from the used open-source 7 *nm* PDK [25]. Each standard cell is characterized for various (7×7) input signal slews and output load capacitances to ensure accurate delay and power analysis later at the circuit level, i.e. when NCFET-aware libraries are used by the static timing analysis and power analysis tool flows. The created libraries are compatible with the existing commercial EDA software (e.g., Synopsys and Cadence) and hence they can be directly deployed. Note that, the NCFET-aware cell library comprises *only* NCFET based cells while the baseline one contains only FinFET based cells. For the sake of completeness, in Fig. 3, we present several validation results obtained from [27], comparing the results of the NCFET model with TCAD results. Fig. 3(a) presents the I_D-V_G characteristics for various ferroelectric layer thicknesses. The larger the thickness, the higher the drain current but with the increase likelihood for hysteresis to appear. Fig. 3(b) depicts the corresponding analysis with respect to the sub-threshold slope (SS) improvement as a function of gate voltage. In Fig. 3(c), we report the obtained improvement in the ON current (I_{ON}) for different ferroelectric thicknesses (t_{fe}) as well as for the baseline ON current represented in Fig. 3 as $t_{fe} = 0$. In Fig. 3(d and e), we show the differential voltage gain (A_V) obtained from the negative capacitance effect as well as the resulting internal voltage gain, respectively. Finally, Fig. 3(f) shows the I_D-V_G characteristics for different gate voltage V_G. As it can be noticed in Fig. 3(a-f), results from our NCFET model come with a very good agreement with TCAD results.

B. NEURAL NETWORK MODELING: FROM CIRCUIT TO SYSTEM LEVEL

Circuit Level: NN inference accelerators consist of large arrays of MADD circuits [4], [5]. The MADD units are pipelined and thus, the overall frequency of the MADD array is mainly defined by the frequency of the individual MADD unit. Analogously, the total energy consumption of the entire MADD array is proportional to the energy consumption of the individual MADD circuit. Therefore, in our analysis,

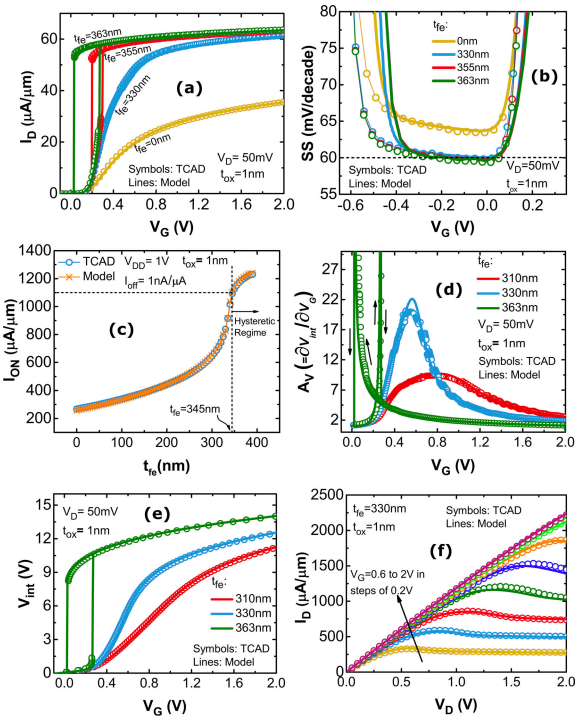

FIGURE 3. Validation results of the employed NCFET model obtained from [27] in which results from the NCFET model for various device characteristics are compared with TCAD results. As can be noticed, a very good matching between our model results and TCAD results is achieved. (a) Shows I_D-V_G results for low V_D at different ferroelectric (t_{fe}) thicknesses. (b) Shows the corresponding sub-threshold slope (SS) improvements. (c) summaries the gain in ON current as a function of t_{fe}. (d and e) report the obtained internal voltage gain from the negative capacitance effect. (f) I_D-V_G results for various gate voltage (V_G). Figures and results are taken from [27].

we consider the frequency and energy consumption of a MADD circuit as a frequency and energy proxy, as typically done in state of the art [19], to assess the frequency and energy improvements that are delivered by NCFET. In our evaluation, we examine varying precision (Q-bit) MADD units. We implement $Q \times Q$ MADD circuits in Verilog RTL using the arithmetic components of the Synopsys DesignWare library (M-2016.12). Each MADD circuit comprises a fixed point multiplier followed by a fixed point adder to accumulate the multiplication results. The size of the multiplier is $Q \times Q$ and the size of the adder is set to 32-bit to avoid accumulation overflow [19]. Hence, note that $Q \times Q$ refers to the arithmetic precision of the MADD unit itself and does not refer to the size of the NN accelerator (e.g., size of systolic MADD array) or the convolution operator. To generate the MADD units, we instantiate two components from the DesignWare library, i.e., a Q-bit multiplier followed by a 32-bit adder. Synopsys Design Compiler is used to synthesize the different MADD circuits and Synopsys PrimeTime is used to perform Static Timing Analysis of the obtained netlists. During synthesis, the "`compile_ultra`" option is used and we instruct Design Compiler to break the hierarchy and flatten the design, in order to obtain well-optimized netlists. Post-synthesis timing simulation is then performed using Mentor QuestaSim to calculate the switching activity of the synthesized netlist. Finally, a power analysis is performed using the extracted switching activity and PrimeTime to accurately calculate the power consumption of each MADD circuit. Every MADD circuit is synthesized and evaluated targeting both the conventional baseline FinFET as well as the NC-FinFET. For our analysis we instantiate fully-optimized arithmetic units from the circuit library and we treat the obtained circuit implementations as black boxes, i.e., we only evaluate their hardware characteristics and we do not interfere with their implementation. Evaluating the efficiency and the optimizations of the synthesis tool is out of the scope of this article.

System Level: We employ an asymmetric min/max post-training quantization method [28] by using a zero-point (ZP) in addition to the scale factor (S). This method allows us to map the min and max values of the float representation to the minimum and maximum range of the quantized one. Once the NN is trained using the default 32-bit float representation, the tensor weight quantization to Q-bit is performed as in:

$$\begin{aligned} x_{quant} &= S * x_{float} - ZP \\ S &= \frac{x_{quant}^{max} - x_{quant}^{min}}{x_{float}^{max} - x_{float}^{min}} \\ ZP &= x_{float}^{max} * S - x_{quant}^{max} \end{aligned} \tag{5}$$

Note that rounding is needed if the result is not an integer.

All the examined NNs (details in Section VII) are developed using Pytorch v1.3 machine learning library [29]. The NNs are trained considering 32-bit floating-point representation and we capture the corresponding inference accuracy on the evaluation datasets. To investigate how the NNs behave under different representations, we employ the previously-described quantization process as Q-bit post-training quantization. For different Q, we quantize the weights, bias, and activations of the 32-bit float representation and we record the inference accuracy of the quantized representation on the same evaluation datasets. Finally, although the quantization method of [28] is employed, our analysis is orthogonal to any quantization approach and similar results are expected.

V. MADD UNIT EVALUATION

A. FREQUENCY AND ENERGY ANALYSIS

In the following, we evaluate the impact of the NC-FinFET on improving the energy and frequency of MADD circuit which is the basic building block of NN inference hardware accelerators. To cover a wide range of scenarios, we consider seven precision levels starting from 6-bit up to 12-bit and we evaluate the frequency and energy consumption of the respective 6×6 to 12×12 MADD circuits. During synthesis, the MADD circuits are implemented targeting both the 7nm FinFET and NC-FinFET libraries. For each circuit, a randomly generated dataset of one million inputs is used to perform post-synthesis timing simulations and then extract the induced switching activity required for accurate power analysis. Leveraging the higher frequency delivered

by NC-FinFET, the operating voltage of NC-FinFET circuits can be scaled down while the baseline frequency (i.e., the maximum frequency provided by the baseline FinFET at 0.7V) can be still sustained. At every reduced voltage, we analyze the efficiency improvement in MADD circuit obtained by NC-FinFET. In our *speed optimization scenario* (details in Section III), we examine the frequency gain obtained by NC-FinFET when MADD circuit consumes the same energy as in the baseline FinFET. In our *energy optimization scenario* (details in Section III), we examine the energy reduction obtained by NC-FinFET when MADD circuit operates at the maximum frequency as in the baseline FinFET. In total, more than 250 MADD configurations (i.e., different Q-bit precision and/or voltage levels and/or operating frequencies) have been evaluated in our experimental results.

Fig. 4 (a and b) presents the frequency and energy gains obtained in the *speed optimization* and *energy optimization* scenarios, respectively. For 6-bit to 12-bit precision levels, Fig. 4a reports the frequency of the NC-FinFET and FinFET based MADD circuits when the former consumes the same energy with the latter. The FinFET based MADD is synthesized and simulated at 0.7V and PrimeTime is used to extract its energy consumption. Then, we scale the operating voltage (from 0.2V to 0.7V) of the respective NC-FinFET MADD and conduct exploration in order to identify the highest voltage value, thus the highest frequency, at which the energy consumption of the NC-FinFET MADD circuit is less or equal to the energy consumption of the respective FinFET MADD. As shown in Fig. 4a, the frequency gain delivered by NC-FinFET is 26.6% on average and ranges from 14.3% for the 6 × 6 MADD to 36.4% for the 12 × 12 MADD. For the widely-used 8 × 8 MADD, NC-FinFET achieves a 23.8% higher frequency for the same energy consumption. Similarly, Fig. 4b demonstrates the energy saving obtained by the NC-FinFET. In this analysis, the targeted frequency of the NC-FinFET based MADD circuits is set to be equal to the maximum frequency of the respective FinFET MADD at 0.7V. Then, we perform an exploration to extract the lowest voltage value, thus the lowest energy consumption, that satisfies this frequency constraint. The energy reduction attained in Fig. 4b is on average 49.7% and ranges from 40.1% up to 55.5%. In the case of the 8 × 8 MADD, NC-FinFET delivers 52% lower energy consumption without any frequency loss. All the configurations extracted by our exploration in Fig. 4a and 4b feature 0.5V and 0.4V voltage values, respectively. This is mainly explained by the fact that the different MADDs exhibit a similar structure and thus the same voltage decrease was required. For the energy optimization, 0.4V was the lowest voltage value that could satisfy the frequency constraint. Similarly, 0.5V is the highest voltage value that could compensate for the increased gate capacitance of the NCFET MADDs and thus satisfy the energy constraint of the speed optimization.

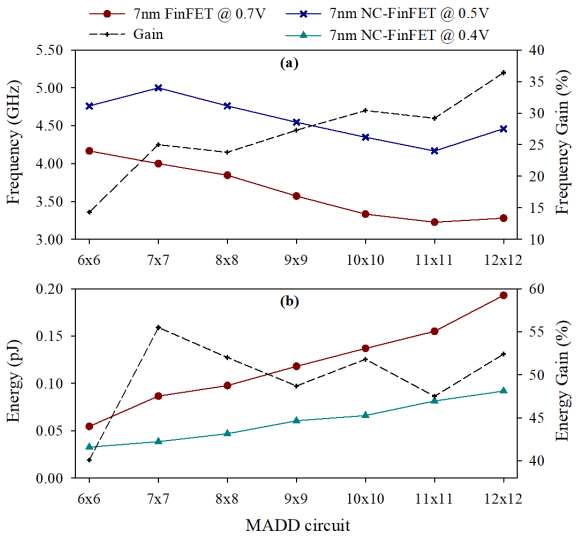

FIGURE 4. Employing NC-FinFET for (a) *speed optimization* and (b) *energy optimization* scenarios at various precision levels in MADD circuits. (a) demonstrates the gain in frequency obtained by NC-FinFET while NC-FinFET MADD circuits still consume the same energy as the baseline FinFET. (b) demonstrates the energy reduction obtained by NC-FinFET, while NC-FinFET MADD circuits still operate at the same frequency as in the baseline FinFET.

B. AREA DISCUSSION

NC-FinFET, in practice, does not result in an area overhead because the transistor footprint remains the same as in the baseline transistor. To realize the effect of negative capacitance, the high-κ layer within the gate stack of the transistor is replaced with a ferroelectric layer. Note that negative capacitance effects can be realized at the same HOf_2 thickness of the baseline transistor (i.e., the same thickness of the baseline high-κ layer of around 2nm) and an increase in the thickness layer results in more gain (i.e., higher internal voltage amplification). However, the length and width of the transistor and hence the transistor area footprint always remain unaffected (i.e., the same).

As mentioned in Section IV-A, the generated NC-FinFET library is the direct counterpart of the baseline FinFET library, i.e., exactly the same cells exist in both libraries. In Fig. 4, the NC-FinFET based MADD units exhibit always lower area than the respective FinFET based MADD ones. This area gain is mainly explained by the fact that due to the internal voltage amplification, NC-FinFET based standard cells are faster [2]. Hence, weaker baseline cells (more area-efficient but slow) become significantly faster in NC-FinFET and thus can be used even under strict synthesis delay constraints. Since we target high performance, the baseline MADD units are synthesized with very strict constraints (i.e., at maximum frequency) and thus weaker cells are mainly not selected by the synthesis tool. However, depending on the optimization scenario examined in Fig. 4, weaker cells can be selected for the NC-FinFET based MADD units. In the speed optimization scenario – in which the NC-FinFET based MADD units are also synthesized under tight timing constraints (i.e., at maximum frequency) – NC-FinFET achieves an average area reduction of 4.1% (up to 6%). In the energy optimization scenario – in which the NC-FinFET based MADD

units are synthesized under relaxed timing constraints (i.e., at the maximum frequency of the respective FinFET based MADD) – the average area reduction is 15.6% (up to 18.7%). In the former case, the area gain is very small and is mainly attributed to the optimizations that the synthesis tool was able to perform (especially in the non-critical paths). In the latter case, the area gain is attributed to the relaxed synthesis constraints and the ability of the tool to select a larger number of weaker, but more area efficient, cells.

VI. A NEURAL PROCESSING UNIT USE CASE

The MADD units examined in Section IV-B are combinational only circuits. However, all other elements (e.g., registers, clock network, etc.) are expected to scale similarly when NC-FInFET is used [2]. To investigate this, we implement a full chip design of the matrix multiply unit (MXU) similarly to the one used in Google TPU [5]. The MXU comprises a large systolic array of 8 × 8 MADD units. We implement the MXU in Verilog RTL and we use Synopsys Design Compiler to synthesize the MXU considering the maximum performance. Then, we use Cadence tool flow to perform the physical implementation of the chip, i.e., GDSII level. Using Cadence Innovus, we design the layout of the floorplan of the chip and the power delivery network (PDN). After that, we perform place and route including clock tree synthesis targeting the maximum performance under the highest optimization constraints. For accurate power and timing analysis, we employ the timing and power signoff tools to report the delay and power of the designed chip after extracting the chip's parasitic. Therefore, we employ Cadence Tempus Timing Signoff tool for delay analysis and Voltus IC Power Integrity signoff tool for power analysis. Moreover, we enabled the on-chip signal and power integrity to consider the impact of the RC-parasitic of the entire chip on delay and power. Finally, for dynamic power analysis, we employ QuestaSim in order to perform timing circuit simulations to extract the switching activity of the final designed chip which is used as input for the power signoff tool for accurate power estimation under real activities. This procedure was done for both the baseline FinFET and NC-FinFET libraries. Note that, the same floorplan, placement, routing, and optimization options are employed for both NCFET and baseline FinFET designs. Finally, for the full chip design of the MXU, we run the aforementioned energy and speed optimization scenarios. For the energy optimization scenario, NC-FinFET achieved 53% lower energy consumption. The respective value, obtained in Fig. 4b for the 8 × 8 MADD is 52%. For the speed optimization scenario, NC-FinFET achieved 22% higher frequency. The respective value, obtained Fig. 4a for the 8 × 8 MADD is 24%. As a result, the gains reported at MADD level are maintained under a full chip design evaluation that considers the entire matrix multiply unit, clock tree, registers, hold and setup times, RC-parasitic, etc. Note that this evaluation targets a specific NN accelerator microarchitecture (i.e., TPU). However, in order to provide a generic evaluation, we focus on investigating the impact of NC-FinFET on the MADD unit that is a core component of any NN accelerator independently of its microarchitecture [4], [5], [17], [19], [30]–[33].

TABLE 1. Neural Network accuracy for varying quantization levels. The Neural Networks are trained on the ImageNet dataset [11] and the asymmetric min/max post-training quantization method [28] is used..

Neural Network	Accuracy (%)							
	6-bit	7-bit	8-bit	9-bit	10-bit	11-bit	12-bit	32-float
ResNet-101	71.49	76.78	77.25	77.26	77.33	77.37	77.37	77.37
SqueezeNet	51.88	56.89	58.07	58.22	58.18	58.17	58.19	58.18
MobileNet	6.49	43.07	62.08	69.9	71.45	71.78	71.87	71.88
ShuffleNet	3.25	41.72	64.29	68.09	69.08	69.19	69.33	69.36
MnasNet	5.09	51.61	66.88	72.49	73.26	73.40	73.45	73.46
AudioCRNN	34.17	53.74	65.41	71.98	74.31	75.07	75.39	75.42
Avg. Accuracy	20.18	49.41	63.35	68.14	69.26	69.52	69.65	69.66

VII. NEURAL NETWORK INFERENCE EVALUATION

In Section V, we demonstrated that NC-FinFET can significantly improve the efficiency of NN inference accelerators in terms of speed and energy consumption. In this section, we perform a system-level evaluation and demonstrate that NC-FinFET can improve both accuracy as well as energy consumption without sacrificing speed. Six NNs are considered in our analysis, i.e., the 101-layer ResNet (ResNet-101) [6], the SqueezeNet v1.1 [7], MobileNet v2 [8], ShuffleNet v2 [9], MnasNet v1.0 [10], and the AudioCRNN [12]. The ResNet-101, SqueezeNet v1.1, MobileNet v2, ShuffleNet v2, and MnasNet v1.0 are image classification NNs and they are trained on the ImageNet dataset [11]. AudioCRNN is an audio classification on and it is trained for the UrbanSound8K dataset [13]. Table 1 reports the accuracy of the examined NNs for different quantization sizes (6-bit to 12-bit) as well as for 32-bit floating point inference. For [6]–[10] the Top-1 accuracy is reported. The last row of Table 1 reports the average accuracy (w.r.t the examined NNs) attained at each quantization size. Hereafter, when referring to the accuracy of a quantization size, we refer to this average value. As shown in Table 1, ResNet-101 and SqueezeNet are amenable to compression and their accuracy is slightly affected by quantization [7], [18]. On the other hand, [8]–[10], [12] are highly impacted by quantization and their accuracy degrades significantly as the quantization size decreases. On average, as shown in Table 1, increasing the quantization size delivers higher accuracy. However, this accuracy improvement comes at the cost of increased hardware requirements. For example, moving from 8-bit to 10-bit quantization increases the inference accuracy by 1.093x (from 63.35% to 69.26%). On the other hand, considering the conventional FinFET (Fig. 4), moving from 8-bit to 10-bit precision MADD units, reduces the frequency by 13% and increases the energy consumption by 40%, respectively. Similarly, targeting speed, moving from 8-bit to 6-bit, decreases the accuracy to only 20.18%.

As shown in Fig. 4, the NC-FinFET based MADD circuits feature significantly higher frequency compared to the respective FinFET ones. Specifically, in Fig. 4a, all the examined NC-FinFET based MADD circuits feature higher frequency than the 6 × 6 FinFET based MADD. Hence,

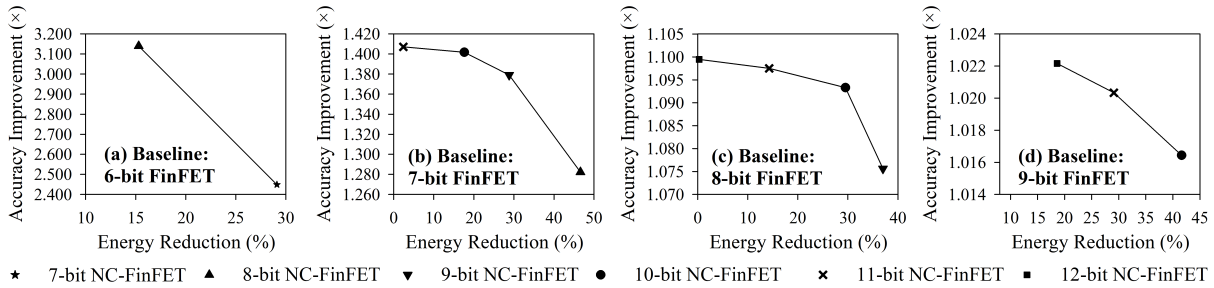

FIGURE 5. The accuracy-energy improvement Pareto-front is delivered by NC-FinFET. In (a)-(d) the accuracy improvement and the energy reduction are reported w.r.t. the FinFET baseline which operates at 0.7V. The accuracy improvement is calculated w.r.t. the average accuracy of each quantization level in Table 1. In (a)-(d) the NC-FinFET circuits operate at the maximum frequency of the respective FinFET baseline, i.e., accuracy-energy improvement without speed loss.

compared to FinFET, NC-FinFET enables increasing the precision of the MADD circuit and thus achieving higher accuracy without any frequency loss. In Fig. 5, we leverage this frequency slack and evaluate the accuracy and energy improvements that can be achieved by exploiting NC-FinFET and adopting higher precision MADD units. For each precision $Q \in [6, 9]$ we perform an exploration to identify all the $Z \times Z$ NC-FinFET based MADD circuits (with $Q < Z \leq 12$) that feature less energy consumption compared to the $Q \times Q$ FinFET based MADD. During the conducted exploration, for each Z value, we set the frequency constraint of the $Z \times Z$ NC-FinFET MADD to be equal to the maximum frequency of the $Q \times Q$ FinFET MADD. Then, we scale the voltage value to extract its lowest value (i.e., lowest energy consumption) that satisfies the frequency constraint. In Fig. 5a-5d, the 6-bit to 9-bit precision FinFET MADD circuits, are used as the baseline and we depict the accuracy-energy improvement Pareto-Front that is obtained by using NC-FinFET based MADD units with higher precision. In Fig. 5a-5d, for the same frequency, the accuracy increases on average by 2.794x, 1.368x, 1.091x, and 1.020x, respectively. In the meantime, in Fig. 5a-5d, the energy consumption decreases on average by 22.2%, 23.9%, 20.3%, and 29.8%, respectively. As illustrated in Fig. 5a, compared to FinFET, NC-FinFET delivers 8-bit accuracy at the speed of 6-bit inference. In other words, compared to 6-bit inference using FinFET, NC-FinFET achieves 3.14x higher accuracy and also 15% energy reduction. Regarding the typically used 8-bit MADD (Fig. 5c), NC-FinFET enables moving up to 12-bit for the same frequency and energy consumption. For example, NC-FinFET enables performing 10-bit inference for the frequency budget of the 8×8 FinFET based MADD circuit. Hence, 1.093x higher accuracy and 30% energy reduction are achieved. Note that in this case, ResNet-101 and SqueezeNet can be still executed at 8-bit quantization since their accuracy increases slightly from 8-bit to 10-bit. Nevertheless, they will be still benefited by the high energy reduction (30%) achieved by the 10×10 NC-FinFET MADD.

NN accelerators can combine multiple MADD units to run the inference with higher precision for the weights and activations and thus, improve the delivered accuracy. For example, a typical accelerator with 8×8 MADD circuits can use two or four MADD units to run the inference with a) a mix of 8-bit activations and 16-bit weights (or vice versa) or b) 16-bits for both the activations and weights. However, in the former case the speed of the accelerator is reduced by half while in the latter the accelerator computes at quarter-speed [5]. Similarly, the energy consumption doubles and quadruples, respectively. As shown in Table 1, the accuracy of the examined NNs barely increases for quantization sizes higher than 10-bits. As aforementioned, NC-FinFET enables using 10-bit precision MADD circuits at the frequency budget of an 8-bit FinFET MADD. As a result, in many cases, when 10-bits for activations and weights deliver satisfying accuracy, we do not need to combine multiple MADD units. Hence, NC-FinFET based NN accelerators can achieve even higher energy savings (more than 30%) and also deliver a significant speed boost. If the 10-bits do not produce satisfying accuracy, the MADD circuits can be still combined, delivering also even higher range for the weights and activations (e.g. 20-bits). *Therefore, NC-FinFET not only improves the energy/speed of the NN inference accelerators but also enables NN developers to rethink their implementations and exploit the higher precision that NC-FinFET delivers to improve the accuracy of their models without trading off for speed and energy*. For example, existing NN architectures trade throughput (e.g., [5] combines many MADD units to enable higher computational precision) or speed (e.g., [30] uses 10-bit MADD units that are 1.15x slower than the 8-bit ones) to achieve higher inference accuracy. Similarly, [19], [32], [33] apply approximations and trade accuracy to improve the speed and/or energy consumption. As a result, NC-FinFET provides new insights and new directions to future NN accelerators architects as well as NN developers.

VIII. NCFET RELIABILITY DISCUSSION

When it comes to reliability, the research study for NCFET is still in its infancy. In the following, we discuss the NCFET reliability for the major reliability degradations: process

variation effects (A), radiation effects (B), and interface traps effects due to aging (C).

A. PROCESS VARIATION EFFECTS

First, we would like to mention that even through the presented analysis in this manuscript is only for the MADD circuit (which is a digital circuit), the characterized NCFET-aware cell library contains a wide range of different standard cell types. In [34], we investigated the impact of various sources of variability on the NCFET transistor compared to the baseline transistor. In addition, the impact of variability effects on the delay of all standard cells is also studied for both baseline and NCFET cells. The following sources of variability were considered: work-function, channel length, Fin height, Fin thickness, effective oxide thickness, ferroelectric thickness, coercive field of ferroelectric and remnant polarization in ferroelectric. In our analysis, we also account for different voltages, in order to explore how reduced voltage in NCFET may impact variation.

Our analysis demonstrated that, at the same operating voltage V_{DD}, NCFET always exhibits less variation ($\sigma I_{ON}/\mu I_{ON}$) compared to the baseline, as shown in Fig. 6. This is due to the internal voltage amplification provided by the negative capacitance effect. When the voltage is reduced, the impact of variation becomes larger in both NCFET and baseline transistors, as expected. However, NCFET suffers less from variability due to the help of negative capacitance

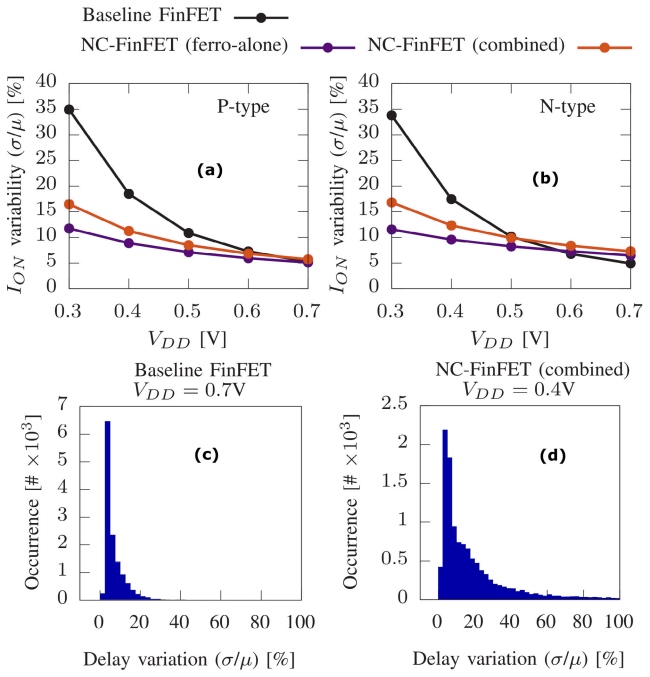

FIGURE 6. Impact of variability on the NCFET compared to the baseline case. (a and b) present the analysis at the device level for p-type and N-type, respectively, w.r.t. the induced variation in the ON current of transistor at different operating voltages. (c and d) present the analysis with respect to the delay variation in all cells within the completed standard cell library. In (c), we present the case of baseline operated at the nominal voltage of 0.7V and in (d), we present the NCFET case operated at 0.4V. Results obtained from [34].

that improves the electrostatic integrity. In other words, the obtained voltage amplification compensate to some degree the reduction in operating voltage. However, when we compare the variation in the baseline transistor that operates at the nominal voltage (i.e., 0.7V) with the NCFET transistor that operates at a lower voltage (e.g., 0.3V and 0.4V), NCFET exhibits a higher variation.

In Fig. 6(a and b), we present the results of variation with respect to ($\sigma I_{ON}/\mu I_{ON}$) for both NCFET and baseline transistors. For the case of NCFET, we show the impact of ferroelectric variability "ferro-alone" (i.e., coercive field of ferroelectric and remnant polarization in ferroelectric), in addition to the "combined" impact of all variability sources together (i.e., work-function, channel length, Fin height, Fin thickness, effective oxide thickness, ferroelectric thickness, coercive field of ferroelectric and remnant polarization in ferroelectric). In Fig. 6(c and d), we show the impact of variability sources on the delay of all standard cells within the standard cell library. In Fig. 6(c), we present the baseline case in which the nominal voltage ($0.7V$) is considered. Whereas, in Fig. 6(d), we present the NCFET case in which a reduced voltage ($0.4V$) is considered.

B. RADIATION EFFECTS

In [35], we have studied the impact of radiation effects on SRAM cells for both baseline and NCFET transistors using calibrated TCAD simulations. It has been demonstrated that the internal voltage amplification obtained in the negative capacitance effect along with the improvement in the electrostatic integrity enables NCFET-based SRAMs to recover faster from a particle strike than the baseline SRAM, i.e., NCFET-based SRAM exhibits a larger critical charge than the baseline SRAM. However, when comparing NCFET-based SRAM (operated at low voltage) with baseline SRAM (operated at higher voltage), NCFET-based SRAM exhibts a lower resiliency to radiation.

C. INTERFACE TRAPS EFFECTS

In [36], we studied the impact of interface traps on the device characteristics of NCFET compared to the baseline transistor using calibrated TCAD simulations. Our analysis showed that, at the same interface trap concentration, the NCFET always exhibits less degradation than the baseline transistor due to the better electrostatic integrity caused by the negative capacitance effect. However, the amplified electric field across the SiO_2 layer within NCFET can lead to a larger interface trap concentration.

IX. CONCLUSION

Negative Capacitance Field-Effect Transistor technology is at the forefront of the steep-slope transistors that overcome the fundamental limit in technology w.r.t voltage scaling. NCFET is rapidly gaining a significant attraction in both academia and industry after it became compatible with the existing CMOS fabrication process. In this work, we are the first to evaluate the impact of steep-slope NCFET transistors

on improving the speed, energy efficiency, and accuracy of Neural Network inference. Through our holistic evaluation, we demonstrated that, compared to conventional 7nm FinFET, NC-FinFET provides up to 36.4% higher frequency and up to 55.5% lower energy. In addition, under the same frequency budget, NC-FinFET enables increasing the precision of the performed computations and thus, increasing the accuracy of the NN inference while still reducing, in the meantime, the energy requirements of the NN accelerators (i.e. no tradeoffs).

ACKNOWLEDGMENT

The work of Hussam Amrouch was done in part at KIT.

REFERENCES

[1] V. V. Zhirnov and R. K. Cavin, "Nanoelectronics: Negative capacitance to the rescue?" *Nature Nanotechnol.*, vol. 3, pp. 77–78, Feb. 2008.

[2] H. Amrouch, G. Pahwa, A. D. Gaidhane, J. Henkel, and Y. S. Chauhan, "Negative capacitance transistor to address the fundamental limitations in technology scaling: Processor performance," *IEEE Access*, vol. 6, pp. 52754–52765, 2018.

[3] Z. Krivokapic, U. Rana, R. Galatage, A. Razavieh, A. Aziz, J. Liu, J. Shi, H. J. Kim, R. Sporer, C. Serrao, A. Busquet, P. Polakowski, J. Muller, W. Kleemeier, A. Jacob, D. Brown, A. Knorr, R. Carter, and S. Banna, "14nm ferroelectric FinFET technology with steep subthreshold slope for ultra low power applications," in *IEDM Tech. Dig.*, Dec. 2017, pp. 15.1.1–15.1.4.

[4] J. Song, Y. Cho, J.-S. Park, J.-W. Jang, S. Lee, J.-H. Song, J.-G. Lee, and I. Kang, "7.1 an 11.5TOPS/W 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile SoC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 130–132.

[5] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, and R. Boyle, "In-datacenter performance analysis of a tensor processing unit," in *Proc. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2017, pp. 1–12.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," 2016, *arXiv:1602.07360*. [Online]. Available: https://arxiv.org/abs/1602.07360

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[9] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[10] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2815–2823.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[12] K. Sanjeevan. (2019). *AudioCRNN, PyTorch Audio Classification: Urban Sounds*. [Online]. Available: https://github.com/ksanjeevan/crnn-audio-classification

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.

[14] S. Mueller, J. Mueller, A. Singh, S. Riedel, J. Sundqvist, U. Schroeder, and T. Mikolajick, "Incipient ferroelectricity in al-doped HfO2 thin films," *Adv. Funct. Mater.*, vol. 22, no. 11, pp. 2412–2417, Jun. 2012.

[15] D. Sarkar, X. Xie, W. Liu, W. Cao, J. Kang, Y. Gong, S. Kraemer, P. M. Ajayan, and K. Banerjee, "A subthermionic tunnel field-effect transistor with an atomically thin channel," *Nature*, vol. 526, no. 7571, pp. 91–95, Oct. 2015.

[16] N. Shukla, A. V. Thathachary, A. Agrawal, H. Paik, A. Aziz, D. G. Schlom, S. K. Gupta, R. Engel-Herbert, and S. Datta, "A steep-slope transistor based on abrupt electronic phase transition," *Nature Commun.*, vol. 6, no. 1, pp. 1–6, Nov. 2015.

[17] G. Desoli, N. Chawla, T. Boesch, S.-P. Singh, E. Guidetti, F. De Ambroggi, T. Majo, P. Zambotti, M. Ayodhyawasi, H. Singh, and N. Aggarwal, "14.1 a 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28nm for intelligent embedded systems," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 238–239.

[18] J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X. Hua, "Quantization networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7300–7308.

[19] P. Gysel, J. Pimentel, M. Motamedi, and S. Ghiasi, "Ristretto: A framework for empirical study of resource-efficient inference in convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5784–5789, Nov. 2018.

[20] (Oct. 2018). *BSIM-CMG Technical Manual*. [Online]. Available: http://www-device.eecs.berkeley.edu/bsim/?page=BSIMCMG

[21] D. Landau and I. M. Khalatnikov, "On the anomalous absorption of sound near a second order phase transition point," in *Doklady Akademii Nauk SSSR*. Pergamon, Turkey: Pergamon, 1954.

[22] K. M. Rabe, M. Dawber, C. Lichtensteiger, C. H. Ahn, and J.-M. Triscone, "Modern physics of ferroelectrics: Essential background," in *Physics of Ferroelectrics*. Berlin, Germany: Springer, 2007, pp. 1–30.

[23] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Lett.*, vol. 8, no. 2, pp. 405–410, Feb. 2008.

[24] (Oct. 2018). *Synopsys EDA Tool Flows*. [Online]. Available: https://www.synopsys.com/

[25] L. T. Clark, V. Vashishtha, L. Shifren, A. Gujja, S. Sinha, B. Cline, C. Ramamurthy, and G. Yeric, "ASAP7: A 7-nm finFET predictive process design kit," *Microelectron. J.*, vol. 53, pp. 105–115, Jul. 2016.

[26] H. Amrouch, S. Salamin, G. Pahwa, A. D. Gaidhane, J. Henkel, and Y. S. Chauhan, "Unveiling the impact of IR-drop on performance gain in NCFET-based processors," *IEEE Trans. Electron Devices*, vol. 66, no. 7, pp. 3215–3223, Jul. 2019.

[27] G. Pahwa, T. Dutta, A. Agarwal, S. Khandelwal, S. Salahuddin, C. Hu, and Y. S. Chauhan, "Analysis and compact modeling of negative capacitance transistor with high ON-current and negative output differential resistance—Part II: Model validation," *IEEE Trans. Electron Devices*, vol. 63, no. 12, pp. 4986–4992, Dec. 2016.

[28] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2704–2713.

[29] *PyTorch v1.3*. Accessed: Dec. 2019. [Online]. Available: https://pytorch.org/docs/1.3.0

[30] W. Qadeer, R. Hameed, O. Shacham, P. Venkatesan, C. Kozyrakis, and M. Horowitz, "Convolution engine: Balancing efficiency and flexibility in specialized computing," *Commun. ACM*, vol. 58, no. 4, pp. 85–93, Mar. 2015, doi: 10.1145/2735841.

[31] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," in *Proc. Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2014, pp. 269–284.

[32] Z.-G. Tasoulas, G. Zervakis, I. Anagnostopoulos, H. Amrouch, and J. Henkel, "Weight-oriented approximation for energy-efficient neural network inference accelerators," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 12, pp. 4670–4683, Dec. 2020.

[33] H. Amrouch, G. Zervakis, S. Salamin, H. Kattan, I. Anagnostopoulos, and J. Henkel, "NPU thermal management," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 11, pp. 3842–3855, Nov. 2020.

[34] H. Amrouch, G. Pahwa, A. D. Gaidhane, C. K. Dabhi, F. Klemme, O. Prakash, and Y. S. Chauhan, "Impact of variability on processor performance in negative capacitance FinFET technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 9, pp. 3127–3137, Sep. 2020.

[35] G. Bajpai, A. Gupta, O. Prakash, G. Pahwa, J. Henkel, Y. S. Chauhan, and H. Amrouch, "Impact of radiation on negative capacitance FinFET," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–5.

[36] O. Prakash, A. Gupta, G. Pahwa, J. Henkel, Y. S. Chauhan, and H. Amrouch, "Impact of interface traps on negative capacitance transistor: Device and circuit reliability," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 1193–1201, 2020.

GEORGIOS ZERVAKIS received the Diploma and the Ph.D. degrees from the Department of Electrical and Computer Engineering (ECE), National Technical University of Athens (NTUA), Greece, in 2012 and 2018, respectively. Before joining KIT, he worked as a primary researcher in several EU-funded projects as a member of the Institute of Communication and Computer Systems (ICCS), Athens, Greece. He is currently a Research Group Leader with the Chair for Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Germany. His research interests include approximate computing, low-power design, design automation, and integration of hardware acceleration in cloud.

IRAKLIS ANAGNOSTOPOULOS (Member, IEEE) received the Ph.D. degree from the Microprocessors and Digital Systems Laboratory, National Technical University of Athens. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, Southern Illinois University, Carbondale. He is also the Director of the Embedded Systems Software Laboratory, which works on run-time resource management of modern and heterogeneous embedded many-core architectures, and he is also affiliated with the Center for Embedded Systems. His research interests include constrained application mapping for many-core systems, design and exploration of heterogeneous platforms, resource contention minimization, and power-aware design of embedded systems.

SAMI SALAMIN (Student Member, IEEE) received the B.Sc. degree in computer systems engineering and the M.Sc. degree (Hons.) from Palestine Polytechnic University, Hebron, Palestine, in 2005 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Chair of Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, under the joint supervision of Prof. Jörg Henkel, and Prof. Hussam Amrouch.

YOGESH S. CHAUHAN (Fellow, IEEE) was with the Semiconductor Research and Development Center, IBM Bangalore, from 2007 to 2010, the Tokyo Institute of Technology in 2010, the University of California Berkeley from 2010 to 2012, and ST Microelectronics from 2003 to 2004. He is currently an Associate Professor with the Indian Institute of Technology Kanpur (IITK), India. He is also the Developer of Industry Standard BSIM-BULK (formerly BSIM6) model for bulk MOSFETs and ASM-HEMT model for GaN HEMTs. His group is also involved in developing compact models for FinFET, nanosheet/gate-all-around FET, FDSOI transistors, negative capacitance FETs, and 2-D FETs. He has published more than 200 articles in international journals and conferences. His research interests include characterization, modeling, and simulation of semiconductor devices. He is also the member of IEEE-EDS Compact Modeling Committee and a fellow of the Indian National Young Academy of Science (INYAS). He received the Ramanujan Fellowship in 2012, the IBM Faculty Award in 2013, the P. K. Kelkar Fellowship in 2015, the CNR Rao Faculty Award, and the Humboldt Fellowship in 2018. He is also the Founding Chairperson of the IEEE Electron Devices Society U.P. chapter and the Vice-Chairperson of IEEE U.P. section. He has served with the Technical Program Committees of IEEE International Electron Devices Meeting (IEDM), IEEE International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), IEEE European Solid-State Device Research Conference (ESSDERC), IEEE Electron Devices Technology and Manufacturing (EDTM), and IEEE International Conference on VLSI Design and International Conference on Embedded Systems. He is also the Editor of IEEE TRANSACTIONS ON ELECTRON DEVICES and the Distinguished Lecturer of the IEEE Electron Devices Society.

JÖRG HENKEL (Fellow, IEEE) received the Diploma and Ph.D. *(summa cum laude)* degrees from the Technical University of Braunschweig. He is currently the Chair Professor of embedded systems with the Karlsruhe Institute of Technology. Before that, he was a Research Staff Member with the NEC Laboratories, Princeton, NJ. His research interest includes co-design for embedded hardware/software systems with respect to power, thermal and reliability aspects. He has received six best paper awards throughout his career from, among others, ICCAD, ESWeek, and DATE. For two consecutive terms, he served as the Editor-in-Chief for the *ACM Transactions on Embedded Computing Systems*. He is also the Editor-in-Chief of the *IEEE Design and Test Magazine*. He is/has been an Associate Editor for major ACM and IEEE journals. He has led several conferences as the General Chair including ICCAD and ESWeek. He serves as the Steering Committee Chair/Member for leading conferences and journals for embedded and cyber-physical systems. He coordinates the DFG Program SPP 1500 ''Dependable Embedded Systems'' and is also a Site Coordinator of the DFG TR89 Collaborative Research Center on ''Invasive Computing''. He is also the Chairman of the IEEE Computer Society and Germany Chapter.

HUSSAM AMROUCH (Member, IEEE) received the Ph.D. degree *(summa cum laude)* from KIT, in 2015. He is currently a Junior Professor heading the Chair of Semiconductor Test and Reliability (STAR) within the Computer Science, Electrical Engineering Faculty, University of Stuttgart and the Research Group Leader with the Karlsruhe Institute of Technology (KIT), Germany. His main research interests include design for reliability and testing from device physics to systems, machine learning, security, approximate computing, and emerging technologies with a special focus on ferroelectric devices. He holds seven HiPEAC Paper Awards and three best paper nominations at top EDA conferences: DAC'16, DAC'17, and DATE'17 for his work on reliability. He is also serving as an Associate Editor at Integration, for TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He has served in the Technical Program Committees for many major EDA conferences, such as DAC, ASP-DAC, ICCAD, and so on. He served as a reviewer in many top journals like IEEE TRANSACTIONS ON ELECTRON DEVICES, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I (TCAS—I), IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, *Transactions on Computers* (TC), and so on. He has around 110 publications in multidisciplinary research areas across the entire computing stack, starting from semiconductor physics to circuit design all the way up to computer-aided design and computer architecture.