FeFET and NCFET for Future Neural Networks: Visions and Opportunities

Mikail Yayla, Kuan-Hsun Chen, Georgios Zervakis, Jörg Henkel, Jian-Jia Chen, Hussam Amrouch
Technical University of Dortmund, Germany
Karlsruhe Institute of Technology, Germany

Citation: https://doi.org/10.23919/DATE51398.2021.9473978

BibTeX:

```
@inproceedings{9473978,
  author={Yayla, Mikail and Chen, Kuan-Hsun and Zervakis, Georgios and Henkel, Jorg and Chen, Jian-Jia a
  booktitle={2021 Design, Automation   Test in Europe Conference   Exhibition (DATE)},
  title={FeFET and NCFET for Future Neural Networks: Visions and Opportunities},
  year={2021},
  volume={},
  number={},
  pages={300-305},
  doi={10.23919/DATE51398.2021.9473978}}
```

FeFET and NCFET for Future Neural Networks: Visions and Opportunities

Mikail Yayla*, Kuan-Hsun Chen*, Georgios Zervakis†, Jörg Henkel†, Jian-Jia Chen*, and Hussam Amrouch‡
*Department of Computer Science, Technische University Dortmund, Germany
†Department of Computer Science, Karlsruhe Institute of Technology, Germany
‡Department of Computer Science, University of Stuttgart, Germany
Email:{mikail.yayla, kuan-hsun.chen}@tu-dortmund.de, {georgios.zervakis, henkel}@kit.edu,
jian-jia.chen@cs.uni-dortmund.de, amrouch@iti.uni-stuttgart.de
(Special Session)

***Abstract*—The goal of this special session paper is to introduce and discuss different emerging technologies for logic circuitry and memory as well as new lightweight architectures for neural networks. We demonstrate how the ever-increasing complexity in Artificial Intelligent (AI) applications, resulting in an immense increase in the computational power, necessitates inevitably employing innovations starting from the underlying devices all the way up to the architectures. Two different promising emerging technologies will be presented: (i) Negative Capacitance Field-Effect Transistor (NCFET) as a new beyond-CMOS technology with advantages for offering low power and/or higher accuracy for neural network inference. (ii) Ferroelectric FET (FeFET) as a novel non-volatile, area-efficient and ultra-low power memory device. In addition, we demonstrate how Binarized Neural Networks (BNNs) offer a promising alternative for traditional Deep Neural Networks (DNNs) due to its lightweight hardware implementation. Finally, we present the challenges from combining FeFET-based NVM with NNs and summarize our perspectives for future NNs and the vital role that emerging technologies may play.**

I. Introduction

Neural Networks (NNs) have been established as the dominant solution in several application domains. Advancements in NNs and DNNs have assisted in boosting the achieved accuracy, resulting in significant improvements in several machine learning applications – especially when it comes to speech and image recognition, gesture detection, and language classification [9]. However, such advancements came at the cost of immense computational demands. In order to achieve the amazing level of accuracy, reached recently in several AI domains, recent DNNs have become deeper and deeper as well as much more complex. As a result, to train such DNNs and/or perform inference rapidly, a massive number of parallel operations (multiplication and addition operations) need to be executed at once exacerbating the computational demands and complexity.

Due to the ever-increasing need to accelerate DNNs inference, targeting to meet tighter and tighter latency constraints, ASIC hardware accelerators have become an integral part of modern systems-on-chip (SoCs). The main computation performed by DNNs is the multiply-accumulate (MAC) operation. DNN accelerators integrate thousands of MAC units to provide a considerable increase in inference speed. For example, Google TPU features 64k MAC units while the embedded oriented Edge TPU comprises 4k MACs. However, performing such a high number MAC operations per cycle results in a significant increase in energy consumption, which might not be tolerated, especially in embedded devices. In addition, recent research demonstrated that integrating such a vast amount of MAC units in a confined area, makes DNN accelerators contingent to high on-chip power densities that rapidly result in excessive on-chip temperatures (see Fig. 1) that form thermal bottlenecks [2]. As a result, DNN accelerators are obliged to sacrifice performance to abide to temperature constraints [2]. The latter further highlights the need for novel solutions that will be able to sustain and/or improve the performance of the DNN accelerators while also satisfying tight power constraints.

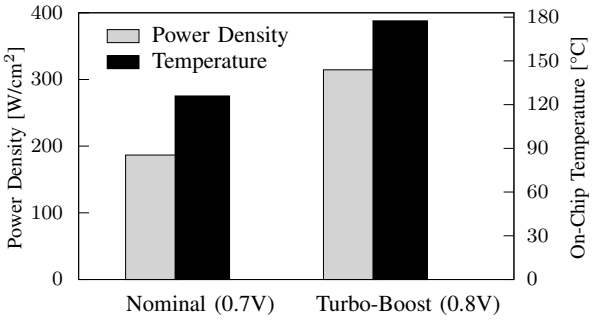


Fig. 1: Power density and corresponding on-hip temperature for a 128×128 systolic MAC array operating at maximum frequency. The maximum heat dissipation that forced-convection air cooling delivers is considered. The $14\,\mathrm{nm}$ FinFET technology node is used for the MAC array implementation, which is done down to the GDSII level. Analysis obtained from [2].

A. Required Innovations for Future Neural Networks

In order to increase the efficiency of NNs and boost the inference speed, while *containing* the computational power demands and hence avoiding thermal bottlenecks, innovations in three main areas are highly required for future NNs. In the following, we summarize those three areas along with an overview of the focus of this paper.

(1) Emerging Non-Volatile Memories: Neural Processing Units, such as Google TPU, include a very large on-chip

memory in order to minimize the need of communication with the off-chip memory. The footprint of such a memory consumes around 30% of the total NPU chip area [9]. On-chip memory is still widely implemented using conventional SRAM cells, which largely suffer leakage power. Apart from the power inefficiency of them, SRAM cells occupy relatively a large area in which every cell consists of 6 transistors. Therefore, the density of SRAM-based on-chip memory is limited.

On the other hand, emerging Non-Volatile Memory (NVM) comes with significantly lower power consumption compared to SRAMs. In addition, NVM offers a much higher density in which a single NVM cell might consist of merely a single transistor. Therefore, replacing conventional SRAM-based on-chip memory in NPUs with NVM-based on-chip memory offers: (a) a large increase in the efficiency due to the much lower power consumption and (b) a higher memory density and hence with the same area footprint of SRAM-based on-chip memory, a larger memory capacity can be embedded using NVM-based memory. This, in turn, allows the NPU to load much larger data for the NNs to be processed and hence reducing the necessity of NPU to communicate with the off-chip memory. In other words, replacing SRAM with NVM brings twofold benefit to the efficiency of NPUs in which not only much lower on-chip power is consumed but also much lower communications with off-chip memory is needed. In this work, we focus on Ferroelectric Field-Effect Transistor (FeFET) as one of the very promising NVM technologies. FeFET is rapidly evolving due to its compatibility to existing CMOS technology. It offers ultra-dense low-power NVM as will be later explained.

(2) Emerging Beyond-CMOS Transistors: Existing CMOS technology suffers from the inability of reducing the operating voltage despite the amazing ability to scale down the transistor feature sizes. This is due to the fundamental limit of sub-threshold swing (SS) of current technology, which is dictated by Boltzmann Tyranny to be always larger than 60mV/dec [15]. Such a fundamental limit imposes strong restrictions on the required operating voltage of circuits and, in fact, it was the key reason behind the discontinuation of Dennard's scaling [1].

Emerging beyond-CMOS technologies that offer steep sub-threshold slope transistors provide a unique opportunity for technology to scale down the voltage while still maintaining the same performance. In this paper, we will focus on Negative Capacitance Field-Effect Transistor (NCFET) as one of the very promising emerging transistors. NCFET technology, similar to FeFET, also employs a ferroelectric layer within the transistor gate stack. Therefore, NCFET is also fully compatible to the existing CMOS fabrication process.

(3) Novel Lightweight NN's Architectures: In most real-world cases, state-of-the-art NNs models rely on a massive number of parameters to achieve high accuracy. Therefore, there is an inherent challenge to run NNs in an efficient way, especially due to the existing memory wall. With the goal of reducing memory accesses and required storage, it has been proposed to employ quantization or even binarization techniques to simplify NN models [7]. Specifically, using Binarized Neural Networks (BNN) significantly reduces the memory needed for parameters and hence makes the computation more efficient, compared to higher precision NNs. Expensive MAC operations can be computed by XNOR followed by bitcount. This enables the possibility to build efficient NN accelerators from simple components, with only a small tradeoff in accuracy. Furthermore, it has been shown that in BNNs, the impacts of bit errors are well-behaved (a flip of one parameter only shifts values by a fixed amount [3]), when compared to floating point NNs (a flip of one bit may cause values to become extremely high). This provides BNNs with an inherent resiliency against errors and make them an excellent candidate to be used with emerging memories, because the latter often offer high energy saving but at the cost of lower reliability. In this work, we explain how FeFET-based memories are subject to errors induced by run-time and design-time variations and demonstrate the high robustness of BNNs against errors, which opens doors for combining BNNs with FeFETs towards energy-efficient, yet reliable NNs.

II. The Role of NCFET Technology in Improving the Efficiency and Accuracy of NNs

Negative Capacitance Field-Effect Transistor (NCFET) technology is at the forefront of emerging beyond-CMOS technologies due to (i) its ability to overcome one of the fundamental limits in existing CMOS technology related to sub-threshold swing of 60mV/dec at the room temperature [1], [15], [20] and (ii) its compatibility with the existing fabrication process of CMOS [11]. The latter is a key for any emerging technology to be adopted by the semiconductor industry as it paves the way for commercial usages with minimal cost overheads.

NCFET aims at increasing the "steepness" of MOSFET transistors towards pushing SS beyond its fundamental limit. This is achieved through *replacing* the traditional high-κ material with a ferroelectric (FE) material. In practice, NCFET dopes the hafnium (HfO_2)-based material – which is widely used in existing CMOS technologies to grow high-κ dielectrics – with zirconium to realize ferroelectricity [11]. The FE layer, under certain conditions of capacitance matching, provides an internal voltage amplification due to the presence of negative capacitance effects. *As a result, the ON current of transistor becomes larger, while the operating voltage remains the same*.

As a matter of fact, the switching speed of any circuit is determined and governed by the ON current (I_{ON}) of the individual transistors that form the critical paths of the circuit. Therefore, employing NCFET enables circuits to achieve a smaller delay and hence be clocked at higher frequency. Alternatively, the same ON current can be achieved but at a lower V_{DD} and without decreasing the threshold voltage V_T, i.e., no increase in the leakage power. Considering the quadratic relation between power and V_{DD}, NCFET delivers significant power saving without performance loss (i.e., no trade-offs).

As mentioned earlier, the core component and basic building block of DNN accelerators is the MAC unit. The overall frequency of the DNN accelerator is mainly defined by the frequency of the individual MAC unit. Analogously, the total power of the entire DNN accelerator is determined by the power consumption of the individual MAC units. Therefore in our

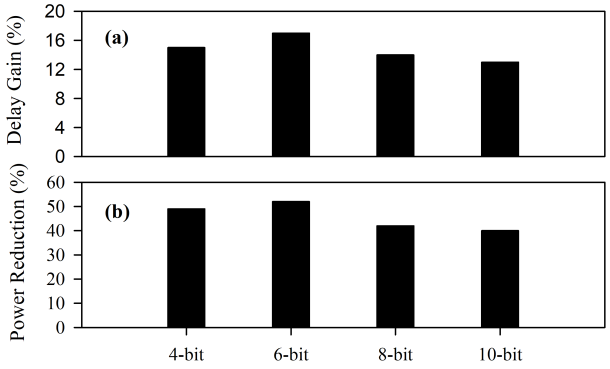


Fig. 2: a) The delay gain achieved by NCFET when the NCFET-based MAC units consume the same energy as the FinFET-based MAC units. b) The power savings delivered by NCFET when the NCFET-based MAC circuits feature the same delay as the FinFET-based ones.

analysis, we consider the MAC unit as our driving circuit to assess the impact of NCFET on future DNN accelerators. Our analysis examines MAC units of varying bit-width (i.e., 4-bit to 10-bit) in order to provide a broad evaluation and cover varying DNN microarchitectures.

In Fig. 2, we evaluate the delay and power improvements that NCFET brings in DNN accelerators. The MAC units are synthesized using Synopsys Design Compiler and then power analyses are performed using Synopsys PrimeTime, using realistic switching activity values. The FinFET-based MAC circuits are mapped to the 14nm baseline FinFET technology library, while the NCFET-based ones are mapped to our 14nm NCFET technology library. The 14nm FinFET and corresponding NCFET libraries were generated as described in [13] and [1], respectively. First, in Fig. 2a we examine the delay gain achieved by NCFET when the NCFET-based MAC units consume the same energy as the FinFET-based MAC units. As shown in Fig. 2a, NCFET achieves 15% lower delay, on average, compared to the baseline FinFET. Note that, this delay gain directly translates to higher throughput and lower overall inference latency. Further details are available in [2]. *Therefore, given an energy optimized DNN accelerator, by employing NCFET, we can significantly improve its performance without breaking the energy optimization.*

In Fig. 2b, we examine the power savings delivered by NCFET when the NCFET-based MAC circuits feature the same delay as the FinFET-based ones. As shown in Fig. 2b, NCFET achieves 46% lower power, on average, compared to the baseline FinFET. Note that, lower power consumption (at the same delay) directly translates to lower energy consumption and also to lower power density, i.e., lower temperature. *Therefore, given a performance optimized DNN accelerator, by employing NCFET, we can significantly decrease its power consumption and satisfy tighter power and temperature constraints.*

Finally, we examine the improvements that NCFET brings at the microarchitecture level. To achieve this, we provide an example that demonstrates how NCFET enables the realization

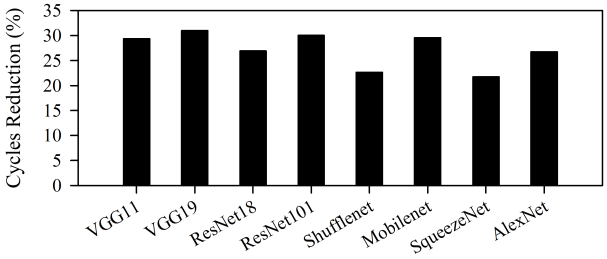


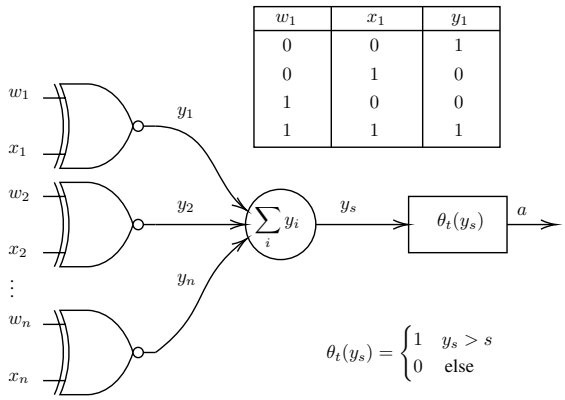
Fig. 3: The inference cycles reduction delivered by the 88×88 NCFET-based 8-bit MAC array compared to the 64×64 FinFET-based one. Both MAC arrays feature the same delay and power consumption.

of more complex DNN accelerators. For this analysis, we consider a microarchitecture similar to the Google Edge TPU. The Edge TPU comprises a 64×64 8-bit systolic MAC array. Exploiting the power gain that NCFET delivers at MAC unit level, for the same total power consumption, we can generate larger MAC arrays that comprise more MAC units. First, based on our analysis in Fig. 2, we performed a full-chip design of a FinFET-based 64×64 MAC array and of a NCFET-based 88×88 one. Both MAC arrays feature the same delay and power consumption (FinFET at 0.7V and NCFET at 0.4V). Note that, 88×88 is selected since it is the largest size that satisfies these requirements when considering 8-bit inference. However, the NCFET-based MAC array features higher throughput since it integrates 1.89x more MAC units. Next, we used the cycle accurate CNN simulator SCALE-Sim from ARM [16] to obtain computation cycles required to run the inference on the aforementioned MAC arrays. Fig. 3 shows the gain in cycles that is obtained by running inference on the 88×88 NCFET-based MAC array compared to running inference on the 64×64 FinFET-based one. Eight state-of-the-art NNs trained on the ImageNet dataset are considered in Fig. 3. On average, for all the examined NNs, NCFET requires 27% less cycles. Note that, the cycle gain directly translates to latency and energy gains since the NCFET and FinFET MAC arrays feature same delay and power consumption. *Therefore, compared to FinFET, for the same power budget, NCFET enables the realization of larger DNN accelerators that deliver significantly higher performance (lower latency) and lower energy consumption.*

In addition, as shown in Fig. 3 this high latency and energy reduction enable running more complex DNNs on resource constrained devices. The latter may eventually translate to higher accuracy. For instance, running ResNet18 on the 88×88 NCFET-based MAC array result to 7.53% higher accuracy, 13% lower latency, and 13% lower energy consumption compared to running the edge-oriented MobileNet on the 64×64 FinFET-based MAC array.

III. The Role of FeFET Technology in Improving the Efficiency of NNs

In the following, we provide a general overview on the FeFET technology and then explain BNNs in detail along with



w_1	x_1	y_1
0	0	1
0	1	0
1	0	0
1	1	1

Fig. 4: A hidden neuron of a BNN, where w_i is the binary weight, x_i the binary input, y_i the XNOR-result, y_s the shifted result, and a the output of the activation function.

discussing why BNNs is an interesting candidate to be used together with FeFETs due to its inherent resiliency to errors.

A. Overview of FeFET Technology

As earlier explained, NCFET replaces the traditional high-κ layer in transistors with a ferroelectric layer. Ferroelectric FET (FeFET) is, in fact, similar but with one key difference regarding the thickness of the included FE layer. When the FE layer thickness becomes relatively high (e.g., 10nm), then negative capacitance effects disappear and instead a *hysteresis-loop* in the electrical characteristics of underlying FET transistor comes into place. The latter is in charge of turning the traditional FET into an NVM device. Compared to other emerging memories, FeFET is considered as one of the most promising technology due to the full CMOS-compatible, as demonstrated in many prototypes by GlobalFoundries [18]. Furthermore, it has been shown that FeFET-based memories is able to provide read and write latencies within 1ns, which narrows the gap with SRAM technology, while FeFET still provides much low power [18]. Another key benefit of FeFET is the high density because every FeFET cell consists of merely one transistor.

One of major drawback of FeFET is the low resiliency to errors. Design-time (caused by manufacturing variability) and run-time (caused by temperature) variations strongly degrade the propriety of FE and hence shrinks the available noise margin. Thus, when reading stored values within FeFETs, errors might appear, unlike SRAMs which are very reliable. Therefore, to deploy FeFETs as on-chip memory, it is necessary to understand and model the underlying probability of error based on the mechanisms by which information is stored.

The key mechanism in FeFET for logic'0' and logic '1' to be stored, is the available dipoles inside the FE. When an electric field is applied, the direction of those dipoles switch (based on the direction on the applied field). When the field is ceased, the dipoles retain their direction. Later, the intensity of provided current by FeFET (i.e., high or low current) based on the dipoles direction enables the sensing circuit to differentiate between logic'0' and logic '1'.

B. What FeFETs Offer to Neural Networks?

Using FeFET memory instead of conventional SRAM on NN accelerators has several key benefits. (1) Non-volatility: During inference, NNs need to access a large number of parameters, and buffer the inputs and the intermediate results. When using FeFET instead of SRAM in NN accelerators, this data will be non-volatile. First, keeping the parameters in FeFET memory without the need to continuously supply energy plays a key role in maximizing the efficiency. Secondly, storing the inputs and intermediate results in FeFET allows for novel computation schemes in systems with energy constraints, such as checkpoint computing, where the device can be powered down after the intermediate results (or inputs) have been stored in FeFET [12]. For a comparison with other emerging NVM technologies, recent studies have found that FeFET-based components in NN accelerators outperform RRAM-based ones [4]. (2) High density: With FeFET, six times more data per area can be stored than in conventional SRAM, due to the difference in memory cell structure. (3) Comparable speed to SRAM: Replacing SRAM in NN accelerators will not lead to a large loss in inference speed. Read and write latencies are within 1ns (close to SRAMs).

C. Which Challenges FeFETs bring to Neural Networks?

Fluctuations in temperature considerably impact the FE properties and can lead to flipping the direction of some dipoles. This manifests itself as changes in the provided current by FeFET in which the sensing circuit will later erroneously decode the stored value. In other words, a bit flip occurs during reading. In [5], [14], we have studied for the first time how temperature increase degrades the electrical properties of FeFETs. Using multi-physics (TCAD) simulations, we modeled how temperature and process variation can cause errors. We demonstrated in [5] that temperature increase non-uniformly degrades the electrical propriety of FeFET when it stores logic '0' or logic '1', leading to asymmetric probability of error.

Therefore, despite the aforementioned benefits of FeFETs, deploying them in NNs necessitates that the NN model to be error tolerant, due to the bit errors caused by temperature and process variation. However, abstracted error models of FeFET are not yet available and research is still in its infancy. Fortunately, several studies have investigated similar scenarios with bit errors using other different memory types. These studies explore the use of approximate memory for NNs, in which reliability (in form of bit errors) is traded for energy efficiency and speed [6], [10].

IV. BNNs for Error-Resilient and Efficient NNs

Among many NN models, the most resource efficient variant are BNNs, which are also highly resilient to bit errors [3]. Since BNNs have binarized weights and activations, the convolution and activation can be computed efficiently by

$$2 * popcount(XNOR(W_i^l, X^{l-1})) - \#bits > s,$$

where *popcount* instruction accumulates the number of bit sets, #bits denotes the number of bits in the XNOR operands, and

s is a learnable threshold parameter, the comparison against which produces a binary activation value [8], [17]. An overview of a BNN neuron operation with XNOR and accumulation (popcount), shift, and threshold for hidden layer neurons is shown in Fig. 4. Using BNNs instead of floating-point NN greatly reduces the needed memory size and accesses, while the costly arithmetic operations can be calculated by computationally efficient bit-wise operations. For instance, considering only binary weight accesses, BNNs require 32 times smaller memory size and 32 times fewer memory accesses compared to floating-point NNs, leading to improved energy-efficiency and speed, while only XNOR, accumulation, and binary thresholding is needed for the majority of BNN operations [8].

A. Error-Tolerance Optimization for BNNs

BNNs can also be trained for error tolerance by bit flip injection during training, as proposed in [6]. Due to the binary properties of BNN neurons, bit flips in one weight or activation only causes a fixed change of the values, for which a proof is provided in [3]. However, if the BNNs are optimized with the standard cross entropy loss and high bit error rates during training, the accuracy degrades significantly. This is undesirable, since using BNNs instead of higher precision variants already necessitates sacrificing a small accuracy loss.

With more sources of bit errors in addition to the FeFET bit errors, e.g. due to approximate components, such as approximate computation units, which are employed in novel NN accelerators [19], the number of bit errors will be even higher. In these cases, using bit flip injections during training with the standard cross entropy loss may lead to even higher decrease of accuracy [3].

Alternative methods need to be investigated to conquer the above challenges. By considering how to train BNNs for bit error tolerance without bit flip injections, we were able to modify a method known from support vector machines for the use of bit error tolerance optimization. By identifying margins in the output layer (as described in [3]), we constructed a modified hinge loss (MHL), which significantly increases bit error tolerance of NNs compared to the state-of-the-art method cross entropy loss (CEL) with bit flip injection during training. If the margins in the NN are not distorted, the MHL can also be applied to quantized or floating-point precision NNs.

In the following, we first present the results showing how the MHL outperforms CEL with bit flip injection. Then, we show how the combination of MHL with bit flip injection increases the bit error tolerance drastically. In the experiments, five BNNs were tested for every plot. The same BNNs are used as in [3]. For BNNs trained with MHL, a parameter search was conducted and the best b was chosen. Here, we present experiment results for evaluating accuracy over bit error rate (BER) from 0% to up to 15% or 30% in Fig. 5 and 6 for the Fashion dataset. For training and evaluation, we use a bit error model which is in line with the assumption in recent studies that use approximate memory [6], [10].

We first compare the MHL alone to CEL with bit flip injection in Fig. 5. In the experiment results, we observe that MHL-trained BNNs have better accuracy over BER than BNNs

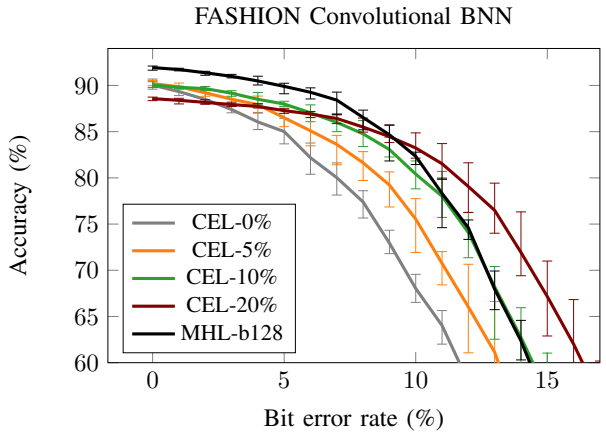


Fig. 5: Accuracy over bit error rate when MHL is used as the loss, compared to CEL with bit flip injections (bit error rates 0%, 5%, etc). The number next to b denotes to which value the parameter b is set to during training with the MHL [3].

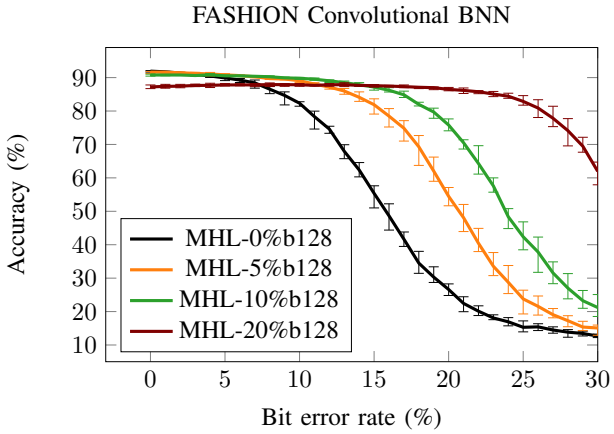


Fig. 6: Accuracy over bit error rate when MHL is combined with bit flip injection (bit error rates 0%, 1%, etc), compared to MHL alone. The number next to b denotes to which value the parameter b is set to during training with the MHL [3].

trained with CEL and bit flip injections, for a BER of up to 10%. From 10% on, the accuracy of CEL-trained BNNs can be better when training with high BERs (e.g. 20%), however, this is after the accuracy drops significantly. Furthermore, the CEL-trained BNNs have significant accuracy degradation for higher BER during training (e.g. CEL-20%), while the MHL-trained BNNs do not suffer from this degradation.

Secondly, we compare the combination of MHL and bit flip injections to the MHL alone. We observe that the accuracy over BER can be drastically improved when using the combination. In Fig. 6, the BER at which the curve drops steeply is extended from 5% (baseline, MHL without bit flips) to 10%, 15%, or 20% BER respectively, depending on the BER during training. There is however a small trade-off in accuracy at 0% BER. With more accuracy trading, higher bit error rates can be tolerated.

Although the MHL increases bit error tolerance, drastic

enhancements can still only be achieved by combining the MHL with bit flip injection during training. On the one hand, bit flip injections introduce overheads during training, which may not scale well for larger NN models and more complex bit error models. Methods to achieve bit error tolerance in NNs without bit flip injection should be continued to be investigated. On the other hand, the combination of MHL and bit flip injection enables the possibility to introduce additional approximate components to the NN inference system, i.e. not only FeFET memory, but also approximate computation units, which increases the number of bit errors in the system.

V. Conclusions and Our Perspectives

(1) Both NCFET and FeFET are promising emerging technologies, which employ ferroelectric material. NCFET increases the efficiency of logic gates, whereas FeFET increases the efficiency of memory. Their compatibility with the existing CMOS fabrication process make them interesting candidates to be adopted by the semiconductor industry.

(2) Building hardware accelerators for DNNs in which (i) SRAM on-chip memory is replaced with FeFET and (ii) conventional FET-based MACs are replaced with NCFET-based MACs opens new doors to significantly accelerate NNs and suppress their computational power demands.

(3) BNN features a lightweight implementation, which makes it a promising candidate to increase the efficiency of NNs. BNNs also exhibit a high resiliency against errors. Therefore, using BNNs together with FeFETs enables NNs to still profit from the high energy savings provided by FeFETs while reliability and accuracy are still maintained despite errors stemming from design-time and run-time variations.

(4) *All in all, we envision that future NNs combine both emerging transistors like NCFETs and emerging NVMs like FeFETs along with lightweight architectures like BNNs towards realizing ultra-low power & ultra-high efficiency AI.*

Acknowledgments

Authors would like to thank Y. Chauhan from IIT Kanpur for NCFET modeling, K. Ni from Rochester Institute of Technology for FeFET modeling, S. Buschjäger and L. Pfahler from TU Dortmund for their support and insights on the optimization of NNs. This work is supported in part by Deutsche Forschungsgemeinshaft (DFG) project OneMemory (project number 405422836) and as part of the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis" (project number 124020371).

References

[1] 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*, 2018.

[2] H. Amrouch, G. Zervakis, S. Salamin, H. Kattan, I. Anagnostopoulos, and J. Henkel. Npu thermal management. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(11):3842–3855, 2020.

[3] S. Buschjäger, J.-J. Chen, K.-H. Chen, M. Günzel, C. Hakert, K. Morik, R. Novkin, L. Pfahler, and M. Yayla. Margin-maximization in binarized neural networks for optimizing bit error tolerance. In *Design, Automation and Test in Europe Conference*, 2021. (accepted).

[4] X. Chen, X. Yin, M. Niemier, and X. S. Hu. Design and optimization of fefet-based crossbars for binary convolution neural networks. In *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1205–1210, 2018.

[5] A. Gupta, K. Ni, O. Prakash, X. S. Hu, and H. Amrouch. Temperature dependence and temperature-aware sensing in ferroelectric fet. In *2020 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–5. IEEE, 2020.

[6] T. Hirtzlin, M. Bocquet, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz. Outstanding Bit Error Tolerance of Resistive RAM-Based Binarized Neural Networks. In *IEEE International Conference on Artificial Intellignence Circuits and Systems (AICAS)*, Hsinshu, Taiwan, Mar. 2019.

[7] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks. In *Advances in neural information processing systems*, pages 4107–4115, 2016.

[8] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 4107–4115. Curran Associates, Inc., 2016.

[9] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 1–12, 2017.

[10] S. Koppula, L. Orosa, A. G. Yağlikçi, R. Azizi, T. Shahroodi, K. Kanellopoulos, and O. Mutlu. Eden: Enabling energy-efficient, high-performance deep neural network inference using approximate dram. In *International Symposium on Microarchitecture*, MICRO '52, 2019.

[11] Z. Krivokapic, U. Rana, R. Galatage, A. Razavieh, A. Aziz, J. Liu, et al. 14nm Ferroelectric FinFET Technology with Steep Subthreshold Slope for Ultra Low Power Applications. In *IEEE Int. Electron Devices Meeting*, Dec 2017.

[12] M. Li, X. Yin, X. S. Hu, and C. Zhuo. Nonvolatile and energy-efficient fefet-based multiplier for energy-harvesting devices. In *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 562–567, 2020.

[13] S. Mishra, H. Amrouch, J. Joe, C. K. Dabhi, K. Thakor, Y. S. Chauhan, J. Henkel, and S. Mahapatra. A simulation study of nbti impact on 14-nm node finfet technology for logic applications: Device degradation to circuit-level interaction. *IEEE Transactions on Electron Devices*, 66(1):271–278, 2018.

[14] K. Ni, A. Gupta, O. Prakash, S. Thomann, X. S. Hu, and H. Amrouch. Impact of extrinsic variation sources on the device-to-device variation in ferroelectric fet. In *2020 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–5. IEEE, 2020.

[15] S. Salahuddin and S. Datta. Use of negative capacitance to provide voltage amplification for low power nanoscale devices. *Nano letters*, 8(2):405–410, 2008.

[16] A. Samajdar, Y. Zhu, P. Whatmough, M. Mattina, and T. Krishna. Scale-sim: Systolic cnn accelerator simulator. *arXiv preprint arXiv:1811.02883*, 2018.

[17] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. How does batch normalization help optimization? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2483–2493. Curran Associates, Inc., 2018.

[18] M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck, et al. A 28nm hkmg super low power embedded nvm technology based on ferroelectric fets. In *2016 IEEE International Electron Devices Meeting (IEDM)*, pages 11–5. IEEE, 2016.

[19] G. Zervakis, H. Amrouch, and J. Henkel. Design automation of approximate circuits with runtime reconfigurable accuracy. *IEEE Access*, 8:53522–53538, 2020.

[20] V. V. Zhirnov and R. K. Cavin. Nanoelectronics: Negative capacitance to the rescue? *Nature Nanotechnology*, 3(2):77–78, 2008.