Tolerating Hard Faults in CNN Accelerators
Supplied at Low Voltage

Yamilka Toca Díaz
yamilka@unizar.es
Dept. of Computer Science and Systems Engineering
Universidad de Zaragoza, Spain

Rubén Gran Tejero
Dept. of Computer Science and Systems Engineering
Universidad de Zaragoza, Spain

Reynier Hernández Palacios
Vicerrectoría de Investigaciones
Universidad de Camagüey, Cuba

Alejandro Valero
Dept. of Computer Science and Systems Engineering
Universidad de Zaragoza, Spain

ABSTRACT
Reducing the supply voltage below the limit of the safe voltage imposed by the worst-case transistor is an effective solution to attain substantial energy savings in current digital circuits. However, this comes at the cost of hard faults appearing in vulnerable transistors, which compromise the correctness of the applications.

This work introduces a pair of low-cost microarchitectural mechanisms to ensure the accuracy of convolutional neural networks running on a specialized accelerator where the supply voltage is set below the limit of the safe voltage.

Experimental results show that, compared to a conventional accelerator design, the proposed techniques reduce the average energy consumption of on-chip memories by 10.5% while maintaining the original (fault-free) accuracy of the networks with a negligible impact on the system performance.

1 INTRODUCTION
Nowadays increasing interest and advances in the field of artificial neural networks is notorious, making it possible to effectively and efficiently model large and complex problems. The concept of neural networks is inspired by the structure and function of the human brain, roughly defined as computational models that allow machines to learn from data and make intelligent decisions. This capability, called machine learning, allows Artificial Intelligence (AI) systems to recognize patterns, make predictions, and continuously improve their performance with experience. There are several types of neural networks, and among them, Convolutional Neural Networks (CNNs) have proven to be particularly effective in processing data related to images and spatial patterns.

To address the computational and memory demands of CNNs, different strategies and technologies have emerged, and one of the most prominent is the use of hardware accelerators designed specifically to run neural networks. Accelerators that speed up the execution of CNNs usually include relatively large and energy-hungry on-chip memories that store either neural network’s activations or weights.

2 PROBLEM DEFINITION
The energy efficiency of current computing systems is compromised by conservative operation guardbands due to variations in the manufacturing process of current CMOS technology nodes. An example is the supply voltage ($V_{dd}$) of the transistor. To ensure a safe system operation against sudden $V_{dd}$ droops, $V_{dd}$ is conservatively set above the limit of the safe voltage ($V_{min}$) imposed by the worst-case transistor. However, significant $V_{dd}$ droops are infrequent events [3]. Moreover, overscaling $V_{dd}$ results in energy wasting, since dynamic energy scales quadratically with the supply voltage.

Many CNN accelerators integrate large and energy-hungry on-chip memories to store network parameters [1, 2]. To reduce energy consumption, a viable solution adopted in various microprocessor components, including on-chip memories, involves relaxing the voltage guardband by lowering $V_{dd}$ toward $V_{min}$ while maintaining a fixed frequency, a technique commonly known as Dynamic Voltage Scaling (DVS). However, aggressively underscaling $V_{dd}$ beyond $V_{min}$ is challenging due to the high number of hard faults appearing in vulnerable bitcells, causing them to remain stuck at a specific logic value [4]. Unfortunately, conventional Error-Correcting Codes (ECC) offer limited coverage to hard faults, requiring larger storage capacities, complex encoders/decoders, and higher energy consumption to ensure a reliable operation [5].

3 MAIN OBJECTIVE
The main objective of this work is to reduce the energy consumption of activation memories in CNN accelerators by aggressively reducing the supply voltage beyond the capability of state-of-the-art proposed solutions, while preserving the accuracy of CNN applications. To address the impact on the accuracy of CNNs of hard faults as a consequence of supply voltage reduction, we have implemented microarchitectural undervoltage mechanisms. For the development of these mechanisms, we have had to perform a thorough study of fault patterns, identifying and analyzing the impact of them on the accuracy of the CNNs. The reader is referred to [6] for further details.

4 CONTRIBUTION
We have designed a pair of low-cost microarchitectural mechanisms to maintain the accuracy of CNN applications in accelerators operating at $V_{dd}$ below $V_{min}$. These mechanisms are transparent to the programmer and do not depend on specific characteristics of CNNs.

The proposed mechanisms enhance the read and write ports of activation memories of CNN accelerators. In particular, they flip the representation of activations to minimize the impact of faults on the activation values. In addition, those activations suffering a high number of faults, where the previous flip strategy fails to minimize
5 CONCLUSIONS AND FUTURE WORK

Aggressively underscaling $V_{dd}$ beyond $V_{min}$ results in a high number of hard faults. This research work proposed a pair of low-cost microarchitectural mechanisms that transform the representation of faulty activation values to minimize the impact on the accuracy of CNNs.

Experimental results have shown that, compared to a conventional CNN accelerator supplied at a safe voltage of 0.6 V, an enhanced accelerator supplied at 0.54 V with our developed mechanisms reduces the average energy consumption of activation memories by 10.5%, while maintaining the original (fault-free) accuracy with a negligible impact on system performance (less than 0.05% for every application).

We are currently exploring new datasets, types of neural networks, and reliability models with a higher number of faults, which will probably require the design of new microarchitectural mechanisms supporting the new conditions of the system.

ACKNOWLEDGMENTS

All authors acknowledge support from grants (1) PID2019-105660RB-C21 and PID2022-136454NB-C22 from Agencia Estatal de Investigación (AEI), Spain, and (2) gaZ: T58_23R research group from Dept. of Science, University and Knowledge Society, Government of Aragon, Spain. The funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES