Ensuring the Accuracy of CNN Accelerators Supplied at Ultra-Low Voltage

Yamilka Toca-Díaz, Rubén Gran Tejero, and Alejandro Valero
Department of Computer Science and Systems Engineering
Aragon Institute for Engineering Research - I3A
Universidad de Zaragoza
Zaragoza, Spain
{yamilka,rgran,alvabre}@unizar.es

***Abstract*—Underscaling the supply voltage (V_{dd}) to ultra-low levels below the safe-operation threshold voltage (V_{min}) brings significant energy savings in digital CMOS circuits but introduces reliability challenges due to increased risk of bitcell permanent faults. This work explores the impact of such faults on the accuracy of a CNN inference accelerator supplying on-chip activation memories at ultra-low V_{dd}. By examining fault patterns, activation values, and memory usage, this paper proposes two microarchitectural techniques exploiting activation outliers and activation memory underutilization. These approaches are cost-effective, do not require programmer intervention, and are application-independent. Experimental results show that the proposed approaches maintain the original CNN accuracy and achieve energy savings by 2.1% and 8.2% compared to the state-of-the-art technique and a conventional accelerator supplied at V_{min}, respectively, with a negligible impact on the system performance (less than 0.25%).**

***Index Terms*—CNN accuracy, deep learning, energy efficiency, permanent faults, ultra-low voltage.**

I. Introduction

Convolutional Neural Networks (CNNs) often rely on specific-purpose hardware accelerators for fast execution. Reducing power consumption in these devices is crucial for energy efficiency and sustainability. However, process variations diminish the energy efficiency of these systems due to conservative operation guardbands. For instance, the transistor's supply voltage (V_{dd}) is set above the safe limit (V_{min}) imposed by the worst-case transistor to mitigate the risk of rare V_{dd} droops, leading to energy wasting as energy consumption quadratically increases with V_{dd}.

CNN inference accelerators usually implement large and energy-hungry on-chip memories consisting of 6T SRAM bitcells vulnerable to process variations. Dynamic Voltage Scaling (DVS) reduces energy by lowering V_{dd} toward V_{min} while maintaining a fixed frequency. Aggressively underscaling V_{dd} below V_{min} saves more energy but implies the appearance of permanent faults in bitcells, requiring complex and energy-hungry Error-Correcting Codes (ECC).

Prior work focusing on permanent faults as a consequence of V_{dd} underscaling in CNN accelerators include V_{dd} adjustments at runtime according to reliability demands [14], FPGA compilation process enhancements [9], or network retraining methods [15]. However, these approaches depend on the programmer or costly application profiling efforts.

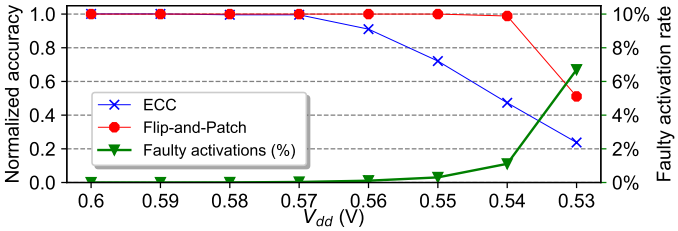

Fig. 1. ECC and Flip-and-Patch (FaP) accuracy varying V_{dd} compared to the golden accuracy. The right Y-axis shows the percentage of faulty activations.

The recent Flip-and-Patch (FaP) technique addresses the impact of permanent faults on the CNN accuracy in two ways. First, 16-bit activations with faults in the most significant byte (M activations) are flipped (i.e., logic values exchange bit positions 15 and 0, 14 and 1, 13 and 2, and so on), minimizing the magnitude deviations caused by faults. Second, a dedicated fault-free patching cache stores a reliable replica of activations with faults in both most and least significant bytes (M&L activations). In contrast, activations with faults in the least significant byte (L activations) are not covered as they do not affect the accuracy. See [12] for further details.

Figure 1 compares the averaged top-1 accuracy of Single-Error Correction Double-Error Detection (SECDED) ECC and FaP for various CNN applications as the V_{dd} of on-chip activation memories of a CNN accelerator reduces below V_{min} (0.6 V). Results are normalized to the golden (fault-free) accuracy. The right Y-axis shows the percentage of faulty activations (M, M&L, and L activations) over the entire on-chip memory of the CNN accelerator. ECC does not hold the accuracy as soon as V_{dd} scales down to 0.56 V. On the other hand, FaP maintains the golden accuracy under a 1.1% faulty activation rate at $V_{dd} = 0.54$ V. However, at 0.53 V, defined as ultra-low V_{dd}, the much higher percentage of faulty activations (6.9%) severely compromises the accuracy of FaP.

Building on FaP, this work introduces Shift-and-Safe (SaS), a pair of microarchitectural mechanisms to handle numerous faults at ultra-low V_{dd}. SaS leverages the presence of activation outliers and the underutilization of activation memories. Like FaP, M activations are turned into L activations with flip operations. Unlike FaP, large value deviations caused by these activations are addressed with a shift-based approach. Finally, instead of a dedicated patching cache, M&L activations are

safely stored in the self activation memory of the accelerator.

Experimental results show that SaS maintains the golden accuracy at ultra-low V_{dd} while reducing average energy consumption by 2.1% compared to FaP. Energy savings increase to 8.2% and 37.7% with respect to a conventional accelerator supplied at V_{min} and nominal voltages, respectively.

II. Background

This section describes the baseline CNN accelerator architecture used in this work. The main components consist of a 16×16 Processing Element (PE) array, on-chip memory storage, dispatchers for every memory, and a control unit [6]. On-chip storage includes a couple of 2 MiB activation memories and a 2 MiB weight memory. The PE array forms a systolic array processor with PEs interconnected through a 2D mesh. Each PE computes 16-bit fixed-point dot-products through partial sums with an activation and a weight. The dataflow in the PE array follows the output stationary approach [10]. This array incorporates intermediate memory buffers to temporarily store and sequentially arrange output activations before forwarding them to the dispatchers.

Like the EIE accelerator [4], activation memories swap input and output roles after the computation of every network layer. These memories are implemented as scratchpad memories, each one including a single read/write port and consisting of eight 256 KiB banks. For simplicity, every layer is stored from address 0x0 onwards and activations are sequentially arranged in memory. Layers exceeding the capacity of activation memories (2 MiB) are spilled to off-chip memory.

Similarly to previous CNN accelerator models [7], network parameters occupy 16 bits and are represented in fixed-point arithmetic, adjusting the number of integer and fractional bits according to the application requirements. Finally, activations are sequentially retrieved from the on-chip memories when an input layer is read, providing 16 consecutive activations (32 bytes) to the dispatchers per memory access. Dispatchers are driven by the control unit, which exploits control information of the current layer to properly feed the PE array.

Finally, like previous work [3], our baseline CNN accelerator has dedicated voltage domains for logic and memory arrays, which allows aggressive voltage underscaling in activation memories while maintaining the remaining hardware components at $V_{dd} \geq V_{min}$ to avoid faults.

III. Proposed Approach: Shift-and-Safe (SaS)

This section discusses the observations that enable SaS, followed by the proposed circuit and its overhead.

A. Shift Technique

This technique is based on a characterization study of activation values. These values usually include outliers, forcing the most significant bits of activations (excluding the sign bit) to be logic '0' in most cases. Particularly, for the studied CNN benchmarks in this work (see Section IV), 99.99%, 99.88%, and 98.96% of the total activations have a '0' in the most significant bit, the two most significant bits, and the three most significant bits, respectively.

Such value distributions present an opportunity to mitigate the influence of faulty bits in L activations and flipped M activations. Specifically, before storing these activations, all the bits except the sign bit are 2-bit left shifted. Subsequently, when retrieving these values from the activation memory, the read operation reverses the shift by two bits to the right, padding the two leftmost magnitude bits with '0'. As a consequence, faulty bitcells affect less significant (two bit positions) activation bits, reducing the impact of these faults on the resulting magnitude.

The optimal shift operation has been experimentally determined. Aggressive shifts involving more than two bits imply larger value losses, whereas a conservative 1-bit shift does not minimize sufficiently the impact of faults on accuracy.

B. Safe-Bank Technique

Shifts are ineffective for M&L activations. The safe-bank approach deals with such activations and it is based on two main observations. First, most activation layers usually demand only a fraction of the available activation memory (e.g., 2 MiB). Particularly, excluding layer sizes larger than 2 MiB, which are spilled to off-chip memory (see Section II), the largest layer size is 1.56 MiB (VGG16) in our studied benchmarks. Unlike the patching cache employed by FaP, which cannot accommodate all the potential M&L activations at ultra-low V_{dd}, the surplus storage at the end of the memory addressing space enables to patch all these activations in such idle memory locations. In this sense, for simplicity, we supply the entire last bank of the activation memory at safe 0.6 V.

The second observation relies on the sequential memory access pattern exhibited by activations (see Section II). Since the order in which activations are consumed is the same in which they are produced, the safe bank is managed as a FIFO queue. When the activation memory acts as output (write) buffer, a pointer keeps track of the next entry to be used in the safe bank. Every new M&L activation is safely stored in that entry and the pointer advances to the subsequent entry. On the contrary, when the activation memory acts as input (read) buffer, the pointer indicates the entry required to restore the following M&L activation. The pointer resets every time the activation memory changes its input/output role.

C. Proposed Circuit

Figure 2 illustrates the required components of the proposed SaS technique in the read port of an activation memory. Remember that a read operation requires to undone transformations in the data representation of stored activations before forwarding them to the dispatcher.

Every stored activation incorporates two control (C) bits to codify the different types of activations. In particular, $C = 00$ identifies reliable activations that do not require any action. $C = 01$ refers to L activations that need to be shifted back. $C = 10$ classifies M activations that necessitate to be flipped back and then shifted back. Finally, $C = 11$ determines M&L activations to be obtained from the safe bank.

Note that C bits are set during post-fabrication testing prior to deploying the accelerator, and they are independent of the

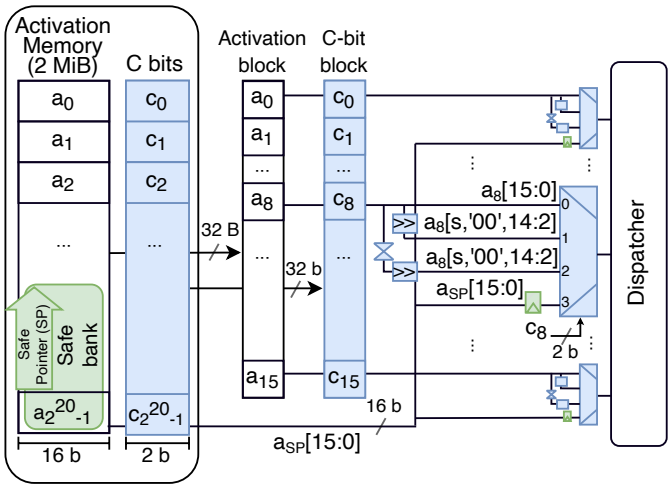

Fig. 2. Proposed Shift (blue) and Safe-bank (green) techniques in the read port of an activation memory.

applications to be run. This process is often employed in practice by traditional error detection/correction techniques [11]. Of course, C bits are supplied at 0.6 V to avoid faults.

After a 32-byte read operation of a block of 16 consecutive activations (e.g., the initial 16 activations in Figure 2) and the corresponding 32 C bits, 16 4-to-1 multiplexers driven by the C bits select among the four types of activations. The figure highlights the required components and subsequent activation bit rearrangements in the selectable inputs of the eighth multiplexer. Reliable activations (input 0) remain unchanged (a_i[15:0]). L activations (input 1) are 2-bit right shifted, preserving the sign (s) bit and introducing two logic '0' in bit positions 14 and 13, resulting in the bit rearrangement a_i[s,'00',14:2]. In the case of M activations (input 2), the flip operation is undone before performing the shift operation, resulting in the same representation as L activations. The management of M&L activations is described next.

When a layer is written, activations identified as M&L ($C = 11$) are sequentially stored one after another in the safe (last) bank, starting at the last memory address and occupying ascending addresses. To do so, we employ a Safe Pointer (SP) stating the next address to be used for this purpose. On the other hand, when a layer is read, the SP pointer is set to the last memory address and M&L activations are read in the same order as they were stored.

For design simplicity, the same memory port is used to access both regular banks and the safe bank. In particular, M&L activations from the safe bank (a_{SP}[15:0]) are read cycle by cycle and temporarily stored in latches at input 3 of the corresponding multiplexers. Once all the M&L activations of a block are ready, the entire block is forwarded to the dispatcher. Finally, note that a similar design is required to incorporate SaS in the write port of the activation memory to properly transform activations.

D. Power, Energy, Area, and Timing Overhead

SaS requires two control (C) bits for every 16-bit activation word. For a 2 MiB activation memory, these bits incur a 256 KiB overhead. While conventional designs already use similar control bits for reliability [11], we conservatively account for their energy, area, and timing overhead. In fact, C bits would not be required for the safe bank. Nevertheless, we have conservatively taken into account their overhead.

TABLE I
POWER AND ENERGY OF DIFFERENT APPROACHES AND SUPPLY VOLTAGES.

	0.6 V	0.54 V	0.53 V	
	DVS	FaP	Base	SaS
Leakage power (mW)	315	296.3	263.8	290
Dynamic read energy (pJ)	83.8	71.3	62.8	69.5
Dynamic write energy (pJ)	66.4	53.7	46.5	52.1

Table I summarizes the leakage power and dynamic energy of an activation memory under different approaches and voltages, obtained with CACTI-P for a 32-nm technology node and ITRS low-power device type [2]. The conventional memory is labeled as DVS (supplied at safe 0.6 V) or Base (powered at ultra-low $V_{dd} = 0.53$ V). FaP and SaS are supplied at 0.54 V and 0.53 V, respectively. SaS overhead includes control bits, shifters, latches, and multiplexers, plus supplying the control bits and safe bank at 0.6 V.

As obtained with CACTI-P, the area of a conventional activation memory is 6.207 mm^2. Adding SaS imposes a 15.4% area overhead. Finally, the proposed read port modifications increase the access time from 2.69 ns to 2.75 ns. We assume this small latency overhead does not compromise the cycle time of the accelerator.

IV. Experimental Evaluation

This section describes the simulation framework used to obtain experimental results, followed by the evaluation of CNN accuracy, system performance, and energy consumption of FaP and the proposed approach.

A. Simulation Environment

We extended TensorFlow 2.5.0 [1] to model our fault-injection framework and the dataflow of the baseline CNN accelerator (see Section II), including the proposed SaS approach. Like previous work [8], we accurately compute memory usage statistics and execution time, assuming a 1 GHz clock frequency and a 3-cycle access latency for on-chip memories [2]. The PE array incurs a 1-cycle penalty for each partial sum and accumulation. Memory statistics and execution cycles are combined with results in Table I to estimate overall energy. Our reliability model is extracted from MoRS configured with undervolted fault map data of a real VC707 Xilinx FPGA [13]. All the results are averaged for ten different fault maps. Finally, four widely used CNN benchmarks are evaluated running a colorectal cancer histology input dataset [5].

B. Impact on Accuracy and System Performance

Figure 3 plots the normalized accuracy of the baseline scheme without any fault protection, FaP, and the proposed SaS technique in activation memories supplied at 0.53 V with respect to a fault-free operation mode with V_{dd} over V_{min}.

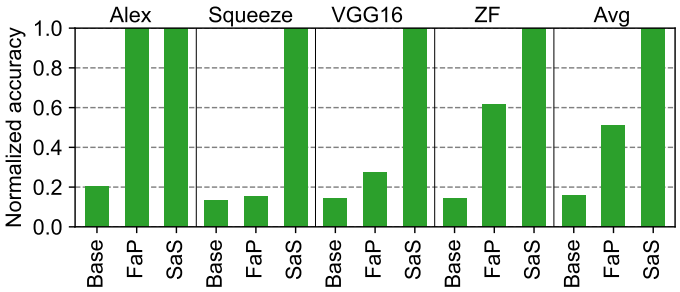

Fig. 3. Normalized accuracy of different approaches operating at 0.53 V with respect to a conventional fault-free operation mode ($V_{dd} \geq V_{min}$).

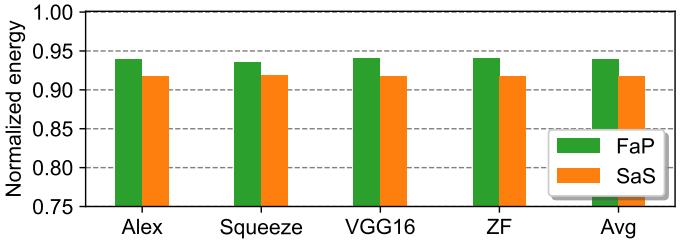

Fig. 4. Normalized energy consumption of an activation memory with FaP (0.54 V) and SaS (0.53 V) compared to DVS (0.6 V).

The baseline scheme severely affects the accuracy in all the benchmarks due to the high number of permanent faults, obtaining a random guessing output in most applications. FaP improves the accuracy, but it is still far from the golden value, cutting the average accuracy in half. Accuracy drops are due to the high number of L activations, flipped M activations, and M&L activations that do not fit into the small patching cache. On the contrary, SaS preserves the golden accuracy by applying shifts to both L and flipped M activations, as well as protecting all the M&L activations in the safe bank.

Managing M&L activations in the safe bank of the proposed design requires additional processor cycles. Compared to a conventional design, the system performance degradation of SaS is by 0.15%, 0.71%, 0.05%, and 0.11% for AlexNet, SqueezeNet, VGG16, and ZFNet, respectively, resulting in a 0.25% average performance loss.

C. Energy Consumption

Figure 4 shows the normalized energy of an activation memory with FaP (0.54 V) and SaS (0.53 V) compared to a conventional memory with DVS (0.6 V). The reported energy comprises both leakage and dynamic expenses, including the overhead of the SaS components (see Section III-D). Energy savings are similar across benchmarks for a given approach. FaP reduces the average energy expenses by 6.1% with respect to DVS. Compared to DVS, SaS obtains average energy savings by 8.2% thanks to reducing V_{dd} down to 0.53 V while maintaining the accuracy. Despite the energy overhead of the SaS components, this technique reduces the average energy consumption by 2.1% with respect to FaP.

These energy savings might seem relatively low. However, it is worth noting that the studied supply voltages just range from 0.6 V (V_{min}) to 0.53 V. Compared to a conventional activation memory supplied at 0.9 V (nominal V_{dd}), the average energy savings of SaS scale up to 37.7%.

V. Conclusions

This work has explored the potential to save energy with supply voltage (V_{dd}) underscaling below the safe voltage (V_{min}) in activation memories of CNN accelerators. It introduces two microarchitectural strategies to address accuracy drops from voltage-induced faults. These approaches are transparent to the programmer and independent of CNN application characteristics. Experimental results have shown that CNN accuracy is preserved with a negligible impact on the system performance, while obtaining energy savings by 2.1% with respect to the state-of-the-art mechanism. Compared to a conventional accelerator supplied at V_{min} and nominal voltages, energy savings are up to 8.2% and 37.7%, respectively.

Acknowledgements

Authors acknowledge support from grants (1) PID2022-136454NB-C22 from *Agencia Estatal de Investigación* (AEI) and (2) gaZ: T58_23R research group from Dept. of Science, University and Knowledge Society, Government of Aragon.

References

[1] M. Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR*, abs/1603.04467, 2016.

[2] R. Balasubramonian et al. CACTI 7: New Tools for Interconnect Exploration in Innovative Off-Chip Memories. *ACM Transactions on Architecture and Code Optimization*, 14(2):1–25, 2017.

[3] A. Chatzidimitriou et al. Assessing the Effects of Low Voltage in Branch Prediction Units. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software*, pages 127–136, 2019.

[4] S. Han et al. EIE: Efficient Inference Engine on Compressed Deep Neural Network. In *Proceedings of the 43rd International Symposium on Computer Architecture*, pages 243–254, 2016.

[5] J. N. Kather et al. Multi-Class Texture Analysis in Colorectal Cancer Histology. *Nature Scientific Reports*, 6, 2016.

[6] N. Landeros Muñoz et al. Gated-CNN: Combating NBTI and HCI aging effects in on-chip activation memories of Convolutional Neural Network accelerators. *Elsevier Journal of Systems Architecture*, 128:1–13, 2022.

[7] J. Lee et al. UNPU: An Energy-Efficient Deep Neural Network Accelerator With Fully Variable Weight Bit Precision. *IEEE Journal of Solid-State Circuits*, 54:173–185, 2019.

[8] L. Mei et al. A Uniform Latency Model for DNN Accelerators with Diverse Architectures and Dataflows. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition*, pages 220–225, 2022.

[9] B. Salami et al. Comprehensive Evaluation of Supply Voltage Underscaling in FPGA on-Chip Memories. In *Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture*, pages 724–736, 2018.

[10] A. Samajdar et al. SCALE-Sim: Systolic CNN Accelerator. *CoRR*, abs/1811.02883, 2018.

[11] J. Tan et al. Combating the Reliability Challenge of GPU Register File at Low Supply Voltage. In *Proceedings of the 25th International Conference on Parallel Architectures and Compilation Techniques*, pages 3–15, 2016.

[12] Y. Toca-Díaz et al. Flip-and-Patch: A fault-tolerant technique for on-chip memories of CNN accelerators at low supply voltage. *Elsevier Microprocessors and Microsystems*, 106:1–13, 2024.

[13] I. E. Yüksel et al. MoRS: An Approximate Fault Modeling Framework for Reduced-Voltage SRAMs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(6):1663–1673, 2022.

[14] J. Zhang et al. ThUnderVolt: Enabling Aggressive Voltage Underscaling and Timing Error Resilience for Energy Efficient Deep Learning Accelerators. In *Proceedings of the 55th ACM/ESDA/IEEE Design Automation Conference*, pages 1–6, 2018.

[15] J. J. Zhang et al. Analyzing and Mitigating the Impact of Permanent Faults on a Systolic Array Based Neural Network Accelerator. In *Proceedings of the IEEE 36th VLSI Test Symposium*, pages 1–6, 2018.