Spiking Neural Network with 3-D IC-based Stacking Memory

NGUYEN Ngo-Doanh - m5262108 Supervised by Prof. DANG Nam Khanh

Abstract

The combination of Neuromorphic Computing and 3D Integrated Circuits - the 3D stacking neuromorphic system can be the most advanced architecture that inherits the benefits of both computing and interconnect paradigms. However, simply shifting to the third dimension cannot exploit the 3D structure and also end up with a low yield rate issue. Therefore, this work proposes a methodology to design 3D stacking synaptic memory for power-efficient operations and yield rate improvement of Neuromorphic Systems (NCs). In this proposed methodology, the synaptic weights are stacked on top of the processing elements, and these weights are split into multiple subsets placed in different layers. Furthermore, with the support of 3D technology, the supply voltage of each layer can be controlled independently which leads to power reduction by scaling down or turning off the supply voltage of the memory layer(s) containing the Least Significant Bits (LSBs) while maintaining acceptable accuracy. On top of that, this work also proposes a methodology to deal with the low-yield rate issue by treating the defective memory cells as noises. In the evaluation with the CMOS 45nm technology, the energy per synaptic operation for MNIST classification, when undervolting two upper memory layers (from 1.1V to 0.8V), reduces by 21.62% while the accuracy only reduces sightly by 0.51%. This energy reduction increases to 66.77% with 6.58% accuracy loss when our system uses both power-gating and undervolting for all memory layers. Furthermore, the system can also improve the yield rate by 0.18% or 12.4% while suffering 0.38% or 1.7% of accuracy loss, respectively.

1 Introduction

Edge devices embedding Artificial Intelligence (AI) have been an emerging computing paradigm recently [1]. However, embedding AI functions into these devices has a lot of challenges because of their resource intensity and power-hungry. As one of many solutions, Spiking Neural Networks (SNNs) show their potential for lightweight inferences compared to other neural network models [2]-[4]. Because, as a mimic of the biological brain, SNNs only transmit information using a sequence of spikes that are believed to be spatial and temporal sparse, which allows them to reduce energy significantly. Moreover, the computation involved in SNNs, especially with Integrate-and-Fire-like models, is comparatively simpler than the conventional neuronal network models. As a result, it reduces the power consumption and hardware area cost.

To exploit the great potential of SNNs, many researchers have investigated deploying these Neuromorphic Computing (NC) systems in recent years. These systems are usually implemented in specific hardware, such as Application-Specific Integrated Circuits (ASICs) or Field-Programmable Gate Arrays (FPGAs), to optimize power and area efficiency, and to perform computations in parallel. In practice, these neuromorphic systems have three main design approaches, which are: (1) 2D-ICs based digital hardware [3], [4]; (2) 2D-ICs based analog mixed-signal hardware [2], [5]; and (3) 3D-ICs based hardware [6], [7]. Nevertheless, as the era of Moore's Law for a single monolithic die nears its end, hardware architectures, particularly memory architectures, are undergoing a transition towards 3D packages or 3D-ICs in order to enhance performance. The architecture of SNNs follows this trend as well [8].

On the other hand, with 3D-IC technologies, memories can be stacked to reduce the hardware footprint. However, we realize that instead of stacking memory banks, we can split the memory words and stack them above each other. In this case, each layer in 3D memory will represent different levels of precision for synaptic weights, such as one, two, or multiple-bit precision. Consequently, the neuromorphic system can selectively deactivate the power supply of individual memory layers that contain the Least Significant Bits (LSBs) in order to conserve energy while still maintaining an acceptable level of accuracy. This is feasible because the absence of LSBs can be treated as a form of noise, and SNNs exhibit resistance to this type of fixed-pattern noise [9]. Based on this feature, this paper presents a novel *in-situ* dynamic quantization hardware architecture of a spiking computing processor using 3D-IC-based stacking memory. In the previous works [10], [11], we have designed a 2D-SRAM-based neuromorphic core connected via 3D-Network-on-Chip, where the memory and the logic computations are placed at the same silicon layer. Based on our experiment, we found out that power consumption of the memory access occupies the major part of the whole system. With our previous architectures, it is difficult to isolate and optimize the power consumption of memory to reduce the overall power consumption of the system. Therefore, in this work, we present a new approach to dynamically reduce the power consumption of memory access with 3D-IC-based

stacking memory and in-situ quantization. The main contributions of this work are summarized in the following:

- A novel low-power methodology to implement neuromorphic architectures with 3D stacking synaptic memory, where the memory word is split into multiple subsets and placed in separate layers.
- With 3D-IC technologies, the under-voltage technique is applied separately to each memory layer in 3D architecture based on the significant bits of synaptic weights. It aims to reduce overall power consumption with acceptable accuracy.
- Consequently, an *in-situ* dynamic quantization for synaptic weight is implemented in this work as the next level of undervolting. The weights are configured in the design phase and stay unchanged during inference. Therefore, the bit precision of synaptic weights is dynamically modified by removing completely the supply voltage of memory layer(s).
- A novel stacking memory mechanism which helps improve the yield rates by accepting imperfection at the top layers.

The rest of this document is organized as follows. Section 2 presents the related works. Section 3 introduces the methodology for 3D-IC implementation. The hardware architecture is shown in Section 4. In Section 5, the performance and power consumption of our spiking computing core in each supply voltage scenario is evaluated. Finally, we end the document with conclusions in Section 6.

2 Related Works

2.1 Low-power Techniques

The *voltage scaling technique* is one of the famous techniques that are widely used for low-power systems. In fact, previous works proved that by applying the under-voltage technique power consumption related to memory could be greatly reduced. For example, Salami *et al.* [12] reduces power consumption by 39% on FPGA on-chip memories, Leng *et al.* [13] saves 20% of power in GPUs, and power consumption of DRAMs in [14] is dropped by 16%. In addition, Minerva [15] lowers the supply voltages of SRAMs to save a total of 2.7x power consumption. In order to accomplish the voltage transformation, the system is required to have an off-chip voltage regulator (VR) with a power switching technique [16], [17] or an on-chip one (i.e.: low-dropout VR [18], [19], switched capacitor VR [20], [21]. Moreover, the under-voltage technique could also be applied to internal components of FPGAs [22] or HBMs (High Bandwidth Memory) [23] to gain around 3x and 2.3x power efficiency, respectively. However, due to the supply voltage reduction, the noise margin of a memory cell is also reduced, which leads to an increase in the probability of errors such as read stability failure, write stability failure, or access time failure [24]. As a result, such small errors could lead to a huge impact on the accuracy of conventional 2D neural network architectures [22]. It is because there is a chance that the MSBs of weights are affected by reducing the supply voltages of SRAMs. However, with 3D technology, the weights can be split into multiple subsets placed in separate layers with isolated supply voltage, which is able to protect the memory layers containing MSBs and reduce the supply voltage of memory layers containing LSBs.

2.2 Power-optimal Memories

Another way to improve the power efficiency of memory is to apply new technologies to restructure the memory cells such as *In-Memory Computing* (IMC), and *3D stacking memory*. For instance, the emergence of IMC methods can be divided into analog IMC [25], [26], [27] and digital IMC [28], [29], [30]. Analog IMC may not be suitable for high-precision applications such AI because as it has the disadvantage of low conversion accuracy limited by the low-cost analog-to-digital converters (ADCs), while digital IMC has the advantage of high computational accuracy. Moreover, the analog IMC is also vulnerable to noise caused by temperature, sneak currents, and many other sources of variations [31]. On the other hand, although the digital IMC has robustness and precision, it consumes more power compared to the analog IMC [32]. For the 3D stacking memory in chips, there are several proposed works [33], [34] to shorten the data movements, which reduces power consumption. With a high bandwidth and a large capacity, 3D stacking of SRAMs has drawn attention for being a large cache in CPUs and a large memory in DNN inference accelerators [35], [36]. The data communication between 3D layers can be wired integration using through-silicon vias (TSVs) [33], [34] or a wireless integration using inductive coupling known as ThruChip Interface (TCI) [37]. However, despite these great benefits of 3D stacking technology, the challenge of this approach is that it has a low yield rate and low reliability. In this paper, to tackle one of these problems, we propose a 3D architecture, which is able to improve the yield rate, by accepting defective layers while maintaining tolerable accuracy.

3 Methodology of 3-D Stacking Memory

3.1 Different Important Levels of Bits

Conventionally, all bits are treated as same as each other regardless of their position in the weight. However, we can simply realize that in terms of value, they are definitely not the same. Although spike neural networks application can be noise resilient, flipping bits due to undervolting or power gating still has different impacts on different positions of the bit. Assuming the weight of $n=8$ bit: $W[0:7]= 10101100$ with one signed bit and seven bits

fractional, the differences in values are shown in Table 1. In summary, flipping bit in the LSBs gives a lesser impact on the value of the weight itself.

Table 1: Difference between bit flipping positions

Value	Original	Flipped Bit Position		
		MSB	3rd bit	LSB
Binary	10101100	10101100	10101100	10101100
Float	-0.34375	0.34375	-0.09375	-0.3515625
Diff. (%)	0 (0%)	+0.6875 (+200%)	+0.25 (+72.727%)	+0.007812 (+2.273%)

Motivated by this, this work presents a method to allow power-reduction targeting LSBs. However, we can quickly notice that power-gating or voltage scaling for LSBs is mostly not possible with the native 2D memory architecture. On the other hand, the 3D architecture is different. It provides different power nets to each stacking layer. Therefore, the voltage-scaling and power-gating techniques could be applied to the memory layers consisting of LSBs to reduce power consumption while maintaining acceptable accuracy.

3.2 Dynamic Low-power Memory Structure

In this proposed methodology, the *n*-bit weights are distributed into different memory layers stacked on each other. It could be also treated as a set of subset bits $\{m_0, m_1, ... m_{M-1}\}$ where m_i is the i^{th} subset of synaptic weights and *M* is the number of subsets. In this case, m_0 contains the most significant bits, and m_{M-1} contains the least significant bits. The number of bits in each subset could be different and can be modified during the design phase. The strategy for *in-situ* low-power structure is acquired by the three following modes (I, II, III), which represent the corresponding low-power techniques. We define those three modes for easier mentioning in the explanation and evaluation.

- **Normal power mode**: The neuromorphic systems operate without power-gating or voltage-scaling.
- **Low-power mode I**: Voltage-scaling is applied to the neuromorphic systems.
- **Low-power mode II:** Power-gating is applied to the neuromorphic systems.
- **Low-power mode III:** Both voltage-scaling and power-gating are applied to the neuromorphic systems.

If the system is currently at low-power mode and the **normal power mode** is detected, the system gradually restores the supply voltage to every inactive memory layer. The order will be bottom-up, which starts from MSBs among all inactive bits. One of the drawbacks of splitting memory weights is having smaller memory cells which lead to lower density and high power consumption. However, we could solve this issue by merging multiple adjacent weights into a single memory cell [5], [38]. Moreover, we utilize multiple power rails for every memory layer to change their power supply. Hence, it is the hardware overhead compared to the traditional voltage scaling. However, our hardware architecture is implemented in 3D and every memory layer has the same hardware area. As a result, compared to the implementation in 2D architecture, there is no overhead in hardware footprint. Another concern of this method is that the number of combinations for configuring and deciding low-power mode for each layer is huge. As a result, a standalone optimization algorithm is required to decide the best operating mode in a specific situation. In this work, the decision is based on our experimental experience.

4 Implementation Hardware Architecture

4.1 3-D Stacking Synaptic Memory

The overview hardware architecture for our proposed methodology is shown in Fig. 1. In detail, Fig. 1(a) shows our software-hardware design methodology with the abstract hardware architecture, shown in Fig.1(b), where we split the logic components and the memory components into separated layers. In addition, Fig.1(c) illustrates our synaptic weights' arrangement in each memory layer while Fig.1(d) presents the block diagram of the logic components in our neuromorphic system. For ease of understanding, the hardware architecture is illustrated as a neuromorphic system consisting of *L=16* Leaky Integrate-and-Fire (LIF) neurons with four synaptic memory layers stacking on top. However, the number of neurons and stacking memory layers could be configured during the design phase. In addition, the output of one neuron could either be transferred to the neurons in the same cluster or in other clusters. On the other hand, the input of one neuron signals the crossbar to extract the corresponding synaptic weight from the memory layers via TSV for LIF computations. The memory layers are stacked on top of the logic computational layer and each memory layer contains a subset of synaptic weights. Those synaptic weights could be either updated from a broadcast message via the address decoder or from the internal Spike-Time Dependent Plasticity (STDP) with self-learning and self-updating functions. Moreover, by dividing the synaptic weights into subsets and placing them on different memory layers, our hardware is able to offer the *in-situ* dynamic quantization for synaptic weights with voltage-scaling and power-gating schemes. These techniques are famous and influential to reduce significantly power consumption in low-power systems.

For a better explanation, the sample hardware for the proposed methodology uses the 8-bit synaptic weights. In addition, it has four memory layers (*M=4*)

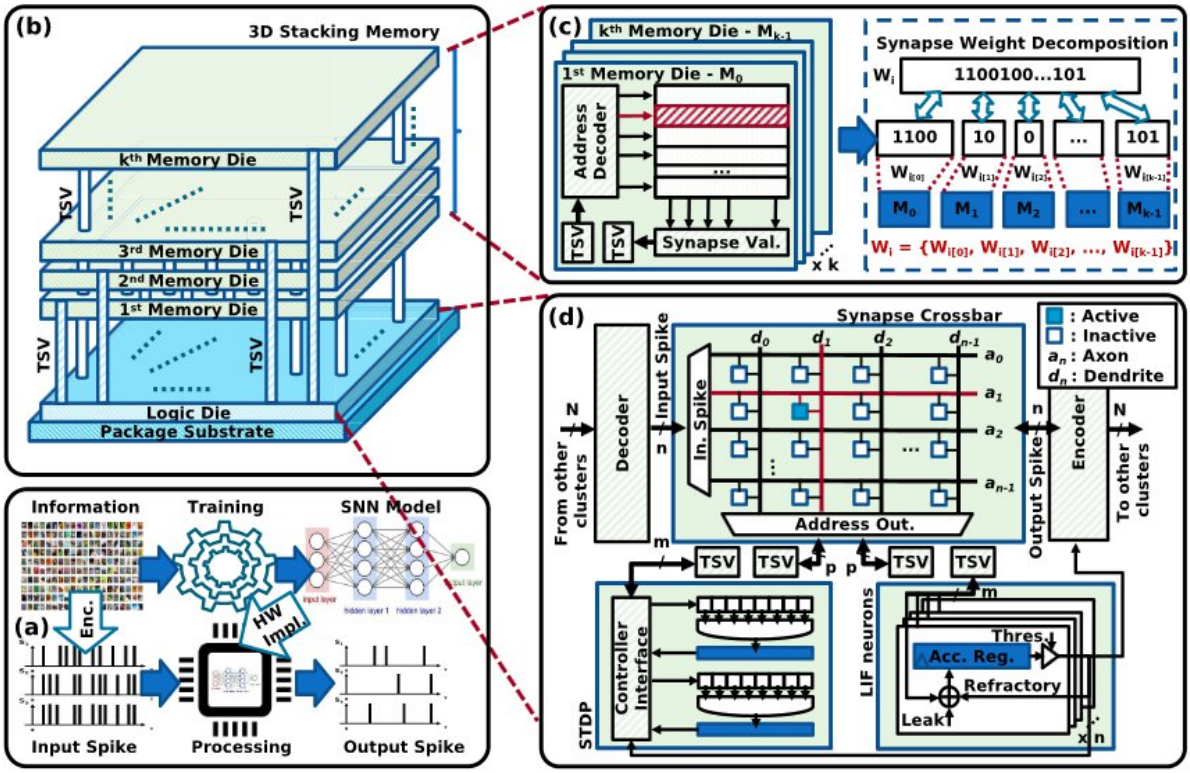

Figure 1: Overview hardware architecture with 3-D stacking memory for the proposal methodology. (a) Overview of SNN hardware implementation. (b) Overview architecture with n stacking memory layers. (c) Synapse weight decomposition with 3-D stacking memory layers. (d) Hardware architecture of each NC core.

stacked on top of PEs and each memory layer contains a 2-bit subset of 8-bit synaptic weights.

The LSBs are placed on the top layer and the MSBs on the bottom layer ({m_0, m_1, m_2, m_3}=W[0:7]). As a result, when applying the voltage-scaling technique or power-gating one to the top memory layer, the power consumption could be reduced from the original power consumption while suffering a small fractional loss in accuracy (flip LSBs). It is only available because of the bit-loss resilience of SNNs [39], where other neural network systems usually drop their accuracy sharply when reducing bit-operation on-fly [22]. In conclusion, there are three benefits to the hardware architecture. First, it takes advantage of 3D implementation, which reduces the transferring distance between memory and PEs. As a result, the power for data transferring can be reduced. Second, the bit-weight quantization could be dynamically activated during the inference without any interruption in system operations. Last, the hardware can partially apply the voltage-scaling technique and the power-gating technique to the memory layer(s), which keeps the MSBs unchanged and only affects LSBs. In addition, LSBs can be reloaded during system operations because the supply voltage is dynamically controlled.

4.2 Power Efficiency with 3-D Stacking Memory

The power consumption of our hardware is similar to other conventional neural network architectures, which is the sum of power consumption by memory storage P_{mem} and power consumption by PEs P_{pe}. In practice, the power consumption from memory is usually dominant, which is about 75% of the total power [40]. It is because the neural network models often require millions of weights to acquire high accuracy and those weights are transferred back and forth in long-distance between memory and PEs. This leads to the huge size of memory, which prolongs the transferring distance and requires more power to transfer those weights in the conventional 2D systems. However, as mentioned above, the 3D design of memory-on-logic brings the two most benefits: distance reduction, and footprint reduction, for neural network models in general, and the SNNs in particular.

On the other hand, the power consumption of CMOS-based circuits could be further expressed as P_{total}, a sum of two components, the dynamic power P_{dyn} (or active power) and the leakage power P_{leak} (or static power).

$$P_{total} = P_{leak} + P_{dyn} \quad (1)$$

Furthermore, those two power consumptions are mathematically represented by the following equations:

$$P_{dyn} = C\, f_{sw}\, V_{DD}^{\,2} \quad (2)$$

$$P_{leak} = K\, N\, I_{leak}\, V_{DD} \quad (3)$$

These equations clearly show that power consumption could be significantly reduced by adjusting the supply voltage. In the case of dynamic

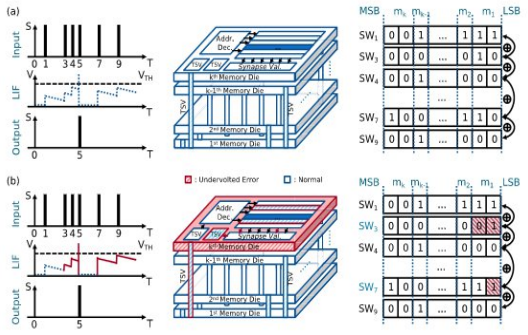

Figure 2: Example of 8-bit synaptic weights' operation with undervolting memory layerse. (a) Operation of our hardware under normal conditions. (b) Operation of our hardware with undervolting for the top memory layer.

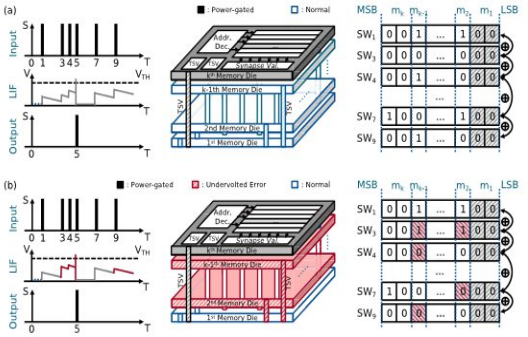

Figure 3: Example of 8-bit synaptic weights' operation with undervolting and power-gating memory layer(s). (a) Operation of our hardware with power-gating the top memory layer. (b) Operation of our hardware with power-gating the top memory layer and undervolting two memory layers.

power, Eq.2 expresses the power reduction in quadratic-fold when scaling down the supply voltage. Moreover, the dynamic power consumption could be further reduced with the power-gating technique, which completely removes the supply voltage. It can only happen in our 3D hardware architecture because of the multiple-layer memory and the noise resilience of SNNs. Likewise, the leakage power consumption is also reduced linearly, as shown in Eq.3, by implementing the same techniques.

To illustrate the power mode I, Fig.2 shows our hardware with undervolting only for the top memory layers and provides the normal supply voltage for the remaining memory layers. In detail, Fig.2(a) shows the normal LIF operation without voltage scaling, and Fig.2(b) demonstrates the LIF operations with the effect of voltage scaling at near-threshold voltage. Here, the red-square areas are the flip-bits due to undervolting. As a result, the flip-bit fault only causes the error in LSBs of synaptic weights and the output spike will not be affected. We first assume that the supply voltage of the top memory layers is reduced by half and there are four stacked memory layers. The total C is *6nF*, $K=1$, the total number of transistors is 10^9, the normal voltage supply is *1.1V*, and the leakage current is $I_{leak}=50pA$. Hence, our hardware, which has a switching frequency of *50MHz*, theoretically could save about *17.92%* power consumption on the memories while the accuracy of our hardware drops insignificantly because of the noise resilience of SNNs. The drop in accuracy will be later evaluated in Section 5. In practice, it could extend approximately the operating time of edge devices by 20%, which is in a power-hungry situation without changing its neural network model and hardware components. Moreover, the accuracy is only trade-offed by a marginal volume.

With the power-gating, our hardware proceeds the *in-situ* synaptic weight quantization by turning the memory layer(s) off if the **low-power mode II** is detected and turning it on if the **normal power mode** is detected. Therefore, the alternation of the total power consumption is from the memory. With $n=8$ as in Fig.1, the expected power reductions are 25% and 50%, for $t=2$ and $t=4$, respectively. Therefore, for each possible value of t, we can define a power-aware mode. In addition, we can also use the voltage-scaling technique for the non-power-gated memory layer(s) to further decrease the overall power consumption. In this case, the system enters the low-power mode III.

Fig.3 shows the example of both **low-power mode II** and **low-power mode III**. With the power-gated top layer, the LSBs of synaptic weights are treated as zeros. It leads to a slight decrease in the value of synaptic weights but our architecture still receives the correct output spike, as shown in Fig.3(a). On the other hand, in the **low-power mode III** (Fig. 3(b)), the synaptic weights in undervolted layers are randomly flipped because of the lack of supply voltage. It also leads to a transformation in the output value of the LIF neuron but the output spike is still correct. It is because the memory layer containing MSBs is untouched. However, the number of untouched MSBs also needs to be considered for the correctness of the SNN model. Despite the noise resilience of SNNs, further dropping the power supply out of the remaining memory layers will cause the spiking computing core to collapse, unable to operate correctly. The evaluation section will demonstrate the experimental results for each operating power-aware mode.

5 Evaluation Results

5.1 Evaluation Methodology

The proposed hardware architecture was implemented in Verilog-HDL, synthesized, and evaluated with commercial CAD tools from Cadence and Synopsys (Cadence Innovus, Synopsys Design Compiler, PrimeTime, Custom Compiler, HSPICE). The physical design of our hardware is implemented with the NANGATE 45-nm library [41] and NCSU FreePDK3D45 TSV [42]. The system memory is 6T SRAM generated from OpenRAM [43] and its BER characteristic, when undervolting is applied, is calculated from Python and is evaluated by HSPICE.

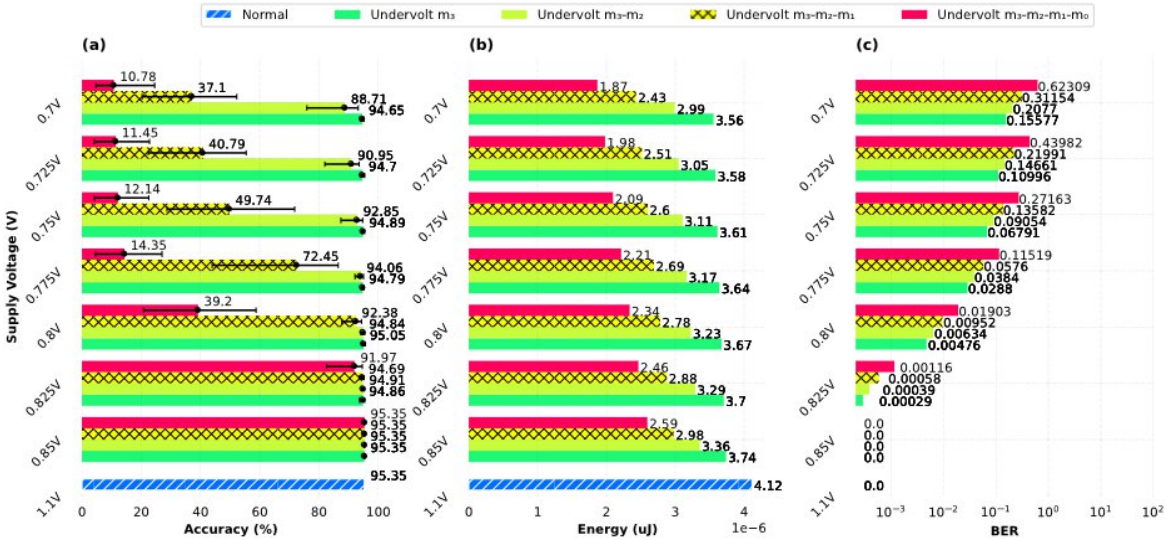

Figure 4: Transformation of BER and accuracy and energy with undervolting memory layer(s). (a) Accuracy when undervolting each combination of memory layer(s). (b) Energy when undervolting each combination of memory layer(s). (c) BER when undervolting each combination of memory layer(s).

In order to evaluate the transformation of power consumption and accuracy, we implemented our hardware as a neuromorphic core with *M=4* memory layers stacked on top of *L=48* LIF modules. The SNN model embedded into the hardware is configured with a neural network of three layers *(784:48:10)* for the MNIST dataset.

We also evaluate the hardware system with the VGG16 model under the CIFAR-10 dataset [44]. Since the hardware design for VGG16 is not available in this work, we estimate the energy consumption via CACTI SRAM's model [45]. The images were encoded into spikes using the rate-coding scheme under the Poisson distribution. In addition, the synaptic weights are trained as *n=8*-bit values for MNIST, and *n=16*-bit values for CIFAR-10. They are split equally into four memory layers of the hardware, which is two bits per layer. Please take note that the configurations of the SNN model and our hardware architecture can also be modified into different ones during the design phase.

First, for the **low-power mode I**, we examine the Signal Noise Margin (SNM) of SRAM cells at near-threshold supply voltages to extract the BER or probability of faults according to materials presented in previous works [46], [47], [48]. The BER is exported through Monte Carlo simulations with PrimeSim HSPICE and mathematical calculation at multiple supply voltages. After that, we insert the faults according to the extracted probabilities into synaptic weights trained from the software model. The position of faults is distributed randomly using the Monte Carlo simulation again with uniform distribution. Because we implement the hardware with four memory layers, the undervolting evaluation is then categorized into four settings. The modified synaptic weights are then loaded into hardware to evaluate the power consumption and the accuracy of the SNN model affected by undervolting.

Second, the transformation of power consumption and accuracy at **low-power mode II** are evaluated. Similar to the **low-power mode I**, the power-gating hardware also has four settings to inspect. However, the accuracy of our hardware is broken when the supply voltage of the third memory layer is turned off. Therefore, in this paper, the evaluation only covers three settings which are: normal setting without power-gating any layers, power-gating one layer, and power-gating two layers. In this case, our hardware treats the bit values of synaptic weights as zero(s) and uses them to perform LIF computations. Similarly, the switching activities of power-gating hardware are then loaded into Synopsys PrimeTime to extract power consumption. Third, the **low-power mode III** are evaluated. Because of the time-consuming simulation, we only pick one case out of all combinations to evaluate the power-accuracy transformation. Finally, we evaluate the hardware complexity and compare our system with other works.

5.2 Undervolting Hardware (Low-power mode I)

As shown in Fig.4, the evaluation of power transformation and accuracy transformation are taken with supply voltages from *0.7V* to *0.85V* with downing *0.025V* per step. Particularly, Fig.4(a) is the evaluation of accuracy transformation, Fig.4(b) is for energy transformation, and the BER of our SRAM is shown in Fig.4(c). According to the NANGATE 45-nm library [41], the voltage threshold of a transistor is around *0.65V*. As a result, we evaluate the transformation from *0.7V* to *0.85V* to capture the best affective region of SNM in the 6T SRAM. Here, the bit order of synaptic weights, as mentioned in Section 4, is that the memory layer m_0 contains the MSBs and the memory layer m_3 contains the LSBs.

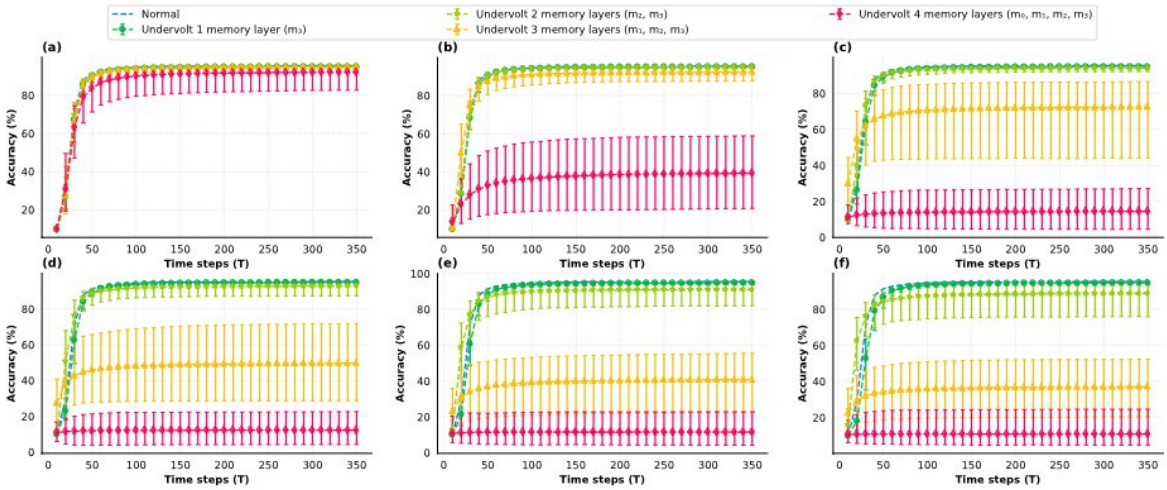

Figure 5: Accuracy with undervolting memory layer(s) in every time step. (a) V_{DD} = 0.825V; BER = 0.00116. (b) V_{DD} = 0.8V; BER = 0.01903. (c) V_{DD} = 0.775; BER = 0.11519. (d) V_{DD} = 0.75V; BER = 0.27163. (e) V_{DD} = 0.725V; BER = 0.43983. (f) V_{DD} = 0.7V; BER = 0.62309.

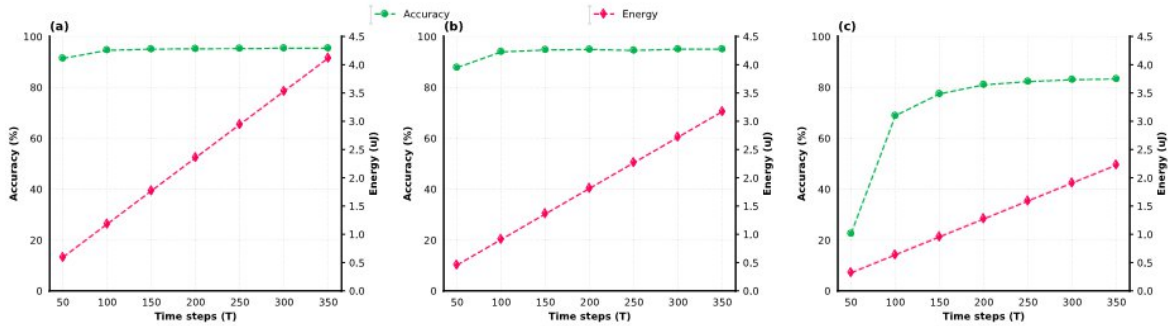

Figure 6: Accuracy and energy consumption of our hardware in different power-gating modes. (a) Trade-off accuracy versus energy at normal operations (no power-gating). (b) Trade-off accuracy versus energy when power-gating m_3. (c) Trade-off accuracy versus energy when power-gating m_3, m_2.

Furthermore, we synchronize all four memory layers ({m_0, m_1, m_2, m_3}) with the same supply voltage ($V_{m0} = V_{m1} = V_{m2} = V_{m3} = V_{DD}$). Please take note that the supply voltages could be independent of each memory layer. To illustrate the transformation of accuracy under the voltage-scaling, Fig.5 shows the accuracy of our hardware per time step, up to 350 time steps. As seen in Fig.5, the average accuracy in all four undervolting modes at a supply voltage of *0.825V* is around *92%*.

The noticeable transformation is that the accuracy significantly swings when undervolting all four memory layers. It is because the MSBs of synaptic weights are affected. However, the BER of SRAMs at this supply voltage is low (*0.00116*). Therefore, the number of modified synaptic weights is low and the worst case for accuracy is around *82.58%*. With the supply voltage scaling down, the average accuracy curves of undervolting three memory layers and undervolting all memory layers are steadily dropped, while the ones from undervolting two memory layers and undervolting one memory layer are only changed slightly. Consequently, undervolting memory layers containing LSBs can lead to achieving high energy efficiency while maintaining acceptable accuracy.

5.3 Power-Gating Hardware (Low-power mode II)

As shown in Fig.6, the accuracy of our power-gated hardware at the 350th computing time-step reaches 95.32%, 94.98%, and 83.28% for each power setting, respectively. This is a very strong indicator that we may be able to offer low-power modes in the trade-off of accuracy loss. In fact, at the 100th computing time-step, the accuracy of our system drops to 94.49%, 93.96%, and 68.71% in each power-gating setting. The accuracy of 4-bit synaptic operations (Fig.6(c)) loses about 15% compared to the 8-bit operations (Fig.6(a)). On the other hand, the accuracy is only reduced slightly by 1% (Fig.6(b)). Here, we can observe that power consumption could be also reduced greatly with the right time step while maintaining a reasonable accuracy. In terms of energy, this reduction in computing time-step leads to a reduction in energy per prediction and energy per Synaptic OPeration (SOP). For the total energy consumption per time-step with the same bit-width synaptic operation, it increases from the 50th time-step to the 350th one approximately by 7x fold.

5.4 Undervolting and Power-Gating Hardware (Low-power mode III)

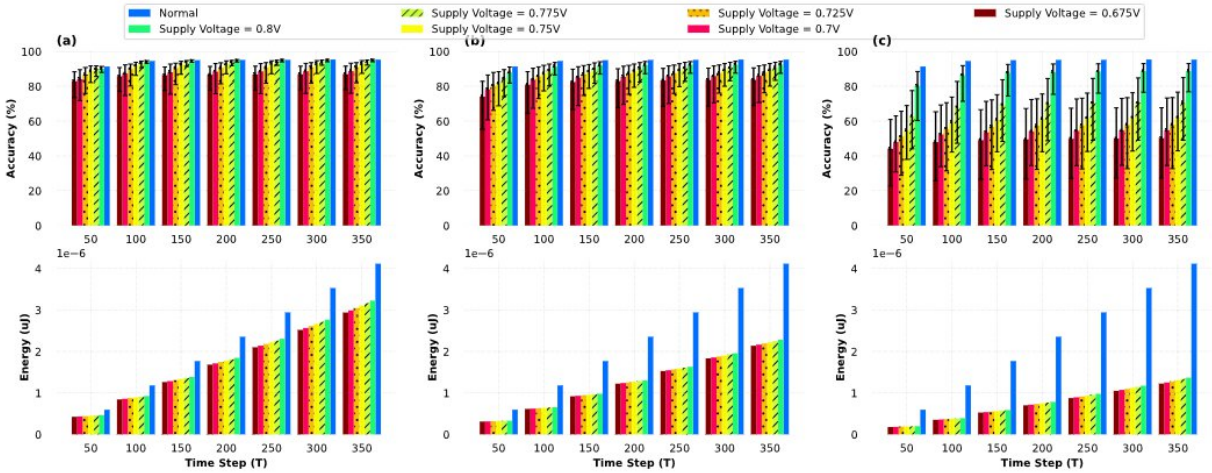

Figure 7: Evaluation of accuracy and energy with both power-gating and undervolting. The supply voltage of the power-gated layer is treated as zero. (a) Accuracy transformation and Energy transformation with undervolting two upper memory layers. (b) Accuracy transformation and energy transformation with power-gating the top memory layer and undervolting two memory layers. (c) Accuracy transformation and energy transformation with power-gating two upper memory layers and undervolting two bottom memory layers.

Table 2: Comparison results between the proposed architecture and existing works

Parameters	TrueNorth [10]	Loihi [1]	ODIN [23]	NASH [68]	Karimi *et al.* [69]	This work					
						Normal Case	Case 1[1]	Case 2[2]	Normal Case	Case 1[1]	Case 3[3]
Benchmark	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST (784:48:10)			CIFAR-10 (VGG16)*		
Accuracy (%)	91.94	96	84	79.4	99.2	95.35	94.84	88.77	91.38	91.26	69.50
Neuron Model	IF	DenMem	LIF & Izhikevicz	LIF	LIF	LIF					
Synaptic Weight Storage	1-bit SRAM	1-to-9-bit SRAM	4-bit SRAM	8-bit SRAM	CTT twin-cell	8-bit SRAM			16-bit SRAM		
Interconnect	2D	2D	2D	3D	2D	3D					
Implementation	Digital	Digital	Digital	Digital	Mix-signal	Digital			Software simulation		
Learning Rule	Un-supervised	On-chip STDP	On-chip Stochastic SDSP	On-chip STDP	Off-chip	Off-chip					
Technology	28nm	14nm FinFET	28nm FD-SOI	45nm	22nm FD-SOI	45nm					
Supply Voltage	0.7-1.05V	0.5-1.2 V	0.55-1 V	1.1 V	0.8 V	0.65V - 1.1V					
Energy per SOP (pJ)	26 (0.775V)	23.6 (0.75V)	8.4	189.3	8	244.28 (1.1V)	191.46[1]	81.16[2]	475.20 (1.1V)	372.13[1]	205.55[3]
Energy per SOP (pJ) (in 14nm)	4.902	23.6	1.078	10.86	4.32	14.02 (1.1V)	10.98[1]	4.65[2]	27.27 (1.1V)	21.35[1]	11.79[3]

[1] Case 1: $\{V_{m_0} = 1.1V; V_{m_1} = 1.1V; V_{m_2} = 0.8V; V_{m_3} = 0.8V\}$ (Low-power Mode I)
[2] Case 2: $\{V_{m_0} = 0.825V; V_{m_1} = 0.8V; V_{m_2} = 0V; V_{m_3} = 0V\}$ (Low-power Mode III)
[3] Case 3: $\{V_{m_0} = 0.825V; V_{m_1} = 0.8V; V_{m_2} = 0.8V; V_{m_3} = 0V\}$ (Low-power Mode III)

As shown in Fig.7(a), the average accuracy of undervolting two memory layerse in 1,000 tests at the supply voltage V_{DD}=*0.8V* is similar to the normal operation of our hardware and this accuracy reduces by 1-2% per undervolting step.

In the worst test, the accuracy drops about 20% compared to the one at the normal operation condition. However, the energy efficiency gains 25%. The energy continues to drop when power-gating is applied to the top layer and undervolting two middle layers (Fig.7(b)). Compared to the normal operation, it is reduced by half yet the average accuracy only reduces slightly. The only noticeable concern is that the range of accuracy is expanded, and the worst accuracy is *55.27%* (dropped about *40%* of accuracy compared to the normal operation). As we continue to drop the supply voltage (Fig. 7(c)), the accuracy swings stronger.

Consequently, the worst accuracy is *22.76%* at V_{m1}=*0.675V* and V_{m0}=0.825V. However, at V_{m1}=0.8V, we can see that the energy is reduced four times compared to the normal operation while reducing *6.57%* in accuracy.

5.5 Comparison

Table 2 represents the comparison results between our work and other existing works [2], [5], [10], [38], [49] which are all based on the MNIST benchmark. In terms of accuracy, the result shows that our system has an accuracy of 95.32% in normal conditions. Furthermore, we pick two other configurations (case 1 and case 2), which use undervolting and power-gating for memory layers. As shown in Table 2, in case 2, with the operation of 4-bit synaptic weights, the accuracy drops by 6.58% compared to the normal operation (8-bit). However, this accuracy is similar to

the works of ODIN [38], which also operates at 4-bit synaptic weight precision.

In terms of power, we compare our work with others using the energy per synaptic operation parameter. Due to the gap in technology, we use the well-known scaling equation from *Stillmaker et al.* [50] to scale down the 14-nm technology node. As shown in Table 2, our hardware consumes 244.28*pJ*, 191.46*pJ*, and 81.16*pJ* at the 45-nm technology node in three cases for 350 time-steps, respectively. After scaling down to the 14-nm technology, our energy per synaptic operation achieves the values, which accordingly are 14.02*pJ*, 10.98*pJ*, and 4.65*pJ*. Furthermore, we also evaluate our methodology with the 16-bit VGG-16 using the CIFAR-10 dataset. As shown in Table 2, the accuracy only drops slightly by *0.12%* while the energy per SOP decreases significantly by *21.68%* in case 1. However, in the case 3, despite the energy reduction of *56.74%*, the accuracy is also reduced seriously by *21.88%*.

In conclusion, these results show that our architecture with 3D stacking memory has an advantage in terms of reducing energy consumption when applying voltage-scaling and power-gating techniques for memory layers. For the MNIST dataset, switching from the normal mode to the low-power mode I, the accuracy drops by *0.51%* to trade-off the energy reduction of *21.62%*. When our hardware switches to the low-power mode III, the accuracy drops by *6.58%* to reduce the energy consumption by *66.77%*. In the case of the CIFAR-10 dataset, with the software simulation, the accuracy also drops by a small fraction (*0.12%*) to reduce *21.68%* energy per synaptic operation when switching from the normal mode to the low-power mode I. Moreover, at the low-power mode III, the accuracy decreases by *21.88%* saving *56.74%* of energy consumption.

6 Conclusion

In this work, we have proposed a methodology to split and stack the synaptic memories for low-power operation. With the 3D technology, the memory can be isolated into different layers, which allows the possibility to separately control the supply voltage of each layer. As a result, the proposed architecture can apply the voltage-scaling technique and also further turn on/off the power supply of one or multiple layer(s) inside it to save the overall energy consumption. In addition, by splitting the synaptic weights into multiple memory layers, the accuracy can be maintained by protecting the memory layer(s) containing the MSBs while dropping the supply voltage of the memory layer(s) containing LSBs. Our future works will extend this work into a very large-scale system using Network-on-Chips with an optimal power-saving strategy.

Reference

[1] M. Merenda, C. Porcaro, and D. Iero, "Edge machine learning for AI-enabled IoT devices: A review," Sensors, vol. 20, no. 9, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/9/2533

[2] M. Davies and et al., "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro, vol. 38, no. 1, pp. 82–99, 2018.

[3] B. V. Benjamin and et al., "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," Proceedings of the IEEE, vol. 102, no. 5, pp. 699–716, 2014.

[4] W. Guo and et al., "Neural coding in spiking neural networks: A comparative study for robust neuromorphic systems," Frontiers in Neuroscience, vol. 15, 2021.

[5] F. Akopyan and et al., "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 10, pp. 1537–1557, 2015.

[6] H. An and et al., "Three-dimensional neuromorphic computing system with two-layer and low-variation memristive synapses," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 3, pp. 400–409, 2022.

[7] G. Orchard and et al., "Efficient neuromorphic signal processing with Loihi 2," in 2021 IEEE Workshop on Signal Processing Systems (SiPS), 2021, pp. 254–259.

[8] A. Ben Abdallah and K. N. Dang, "Toward robust cognitive 3D brain-inspired cross-paradigm system," Frontiers in Neuroscience, vol. 15, 2021.

[9] T. Wunderlich and et al., "Demonstrating advantages of neuromorphic computation: A pilot study," Frontiers in Neuroscience, vol. 13, 2019.

[10] O. M. Ikechukwu, K. N. Dang, and A. B. Abdallah, "On the design of a fault-tolerant scalable three dimensional NoC-based digital neuromorphic system with on-chip learning," IEEE Access, vol. 9, pp. 64 331–64 345, 2021.

[11] K. N. Dang and et al., "MigSpike: A migration based algorithms and architecture for scalable robust neuromorphic systems," IEEE Transactions on Emerging Topics in Computing, vol. 10, no. 2, pp. 602–617, 2022.

[12] B. Salami, O. S. Unsal, and A. C. Kestelman, "Comprehensive evaluation of supply voltage underscaling in FPGA on-chip memories," in Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO), Oct. 2018, pp. 724–736.

[13] J. Leng, A. Buyuktosunoglu, R. Bertran, P. Bose, and V. J. Reddi, "Safe limits on voltage reduction efficiency in GPUs: A direct measurement approach," in Proc. 48th Annu.

IEEE/ACM Int. Symp. Microarchitecture (MICRO), Dec. 2015, pp. 294–307.
[14] K. K. Chang et al., "Understanding reduced-voltage operation in modern DRAM devices: Experimental characterization, analysis, and mechanisms," ACM SIGMETRICS Perform. Eval. Rev., vol. 45, no. 1, pp. 52–51, Jun. 2017.
[15] B. Reagen et al., "Minerva: Enabling low-power, highly-accurate deep neural network accelerators," in Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA), Jun. 2016, pp. 267–278.
[16] L. Di, M. Putic, J. Lach, and B. H. Calhoun, "Power switch characterization for fine-grained dynamic voltage scaling," in Proc. IEEE Int. Conf. Comput. Design, Oct. 2008, pp. 605–611.
[17] Z. Bai, D. Xu, T. Wang, and M.-C. Wong, "A cascaded multilevel battery energy storage based parallel dynamic voltage compensator for medium voltage industrial distribution systems," IEEE Trans. Ind. Informat., early access, Apr. 7, 2023, doi: 10.1109/TII.2023.3265526.
[18] N. Adorni, S. Stanzione, and A. Boni, "A 10-mA LDO with 16-nA IQ and operating from 800-mV supply," IEEE J. Solid-State Circuits, vol. 55, no. 2, pp. 404–413, Feb. 2020.
[19] C.-H. Huang and W.-C. Liao, "A high-performance LDO regulator enabling low-power SoC with voltage scaling approaches," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 28, no. 5, pp. 1141–1149, May 2020.
[20] P. H. McLaughlin, Z. Xia, and J. T. Stauth, "A monolithic resonant switched-capacitor voltage regulator with dual-phase merged-LC resonator," IEEE J. Solid-State Circuits, vol. 55, no. 12, pp. 3179–3188, Dec. 2020.
[21] D. Lutz, P. Renz, and B. Wicht, "12.4 A 10 mW fully integrated 2-to-13 V-input buck-boost SC converter with 81.5% peak efficiency," in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Jan. 2016, pp. 224–225.
[22] B. Salami et al., "An experimental study of reduced-voltage operation in modern FPGAs for neural network acceleration," in Proc. 50th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN), Jun. 2020, pp. 138–149.
[23] S. S. N. Larimi et al., "Understanding power consumption and reliability of high-bandwidth memory with voltage underscaling," 2021, arXiv:2101.00969.
[24] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Statistical design and optimization of SRAM cell for yield enhancement," in Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD), Nov. 2004, pp. 10–13.
[25] E. Lee et al., "A charge-domain scalable-weight in-memory computing macro with dual-SRAM architecture for precision-scalable DNN accelerators," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 68, no. 8, pp. 3305–3316, Aug. 2021.
[26] M. E. Sinangil et al., "A 7-nm compute-in-memory SRAM macro supporting multi-bit input, weight and output and achieving 351 TOPS/W and 372.4 GOPS," IEEE J. Solid-State Circuits, vol. 56, no. 1, pp. 188–198, Jan. 2021.
[27] S. Jain, L. Lin, and M. Alioto, "±CIM SRAM for signed in-memory broad-purpose computing from DSP to neural processing," IEEE J. Solid-State Circuits, vol. 56, no. 10, pp. 2981–2992, Oct. 2021.
[28] H. Kim, Q. Chen, and B. Kim, "A 16K SRAM-based mixed-signal in-memory computing macro featuring voltage-mode accumulator and row-by-row ADC," in Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC), Nov. 2019, pp. 35–36.
[29] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: Enabling in-memory Boolean computations in CMOS static random access memories," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 65, no. 12, pp. 4219–4232, Dec. 2018.
[30] W. Simon, J. Galicia, A. Levisse, M. Zapater, and D. Atienza, "A fast, reliable and wide-voltage-range in-memory computing architecture," in Proc. 56th ACM/IEEE Design Autom. Conf. (DAC), Jun. 2019, pp. 1–6.
[31] M. Hu et al., "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in Proc. 53rd Annu. Design Autom. Conf., Jun. 2016, pp. 1–6.
[32] M. R. H. Rashed, S. K. Jha, and R. Ewetz, "Hybrid analog-digital in-memory computing," in Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD), Nov. 2021, pp. 1–9.
[33] K. Cho et al., "SAINT-S: 3D SRAM stacking solution based on 7nm TSV technology," in IEEE Hot Chips Symp., Aug. 2020, pp. 1–13.
[34] N.-D. Nguyen, X.-T. Tran, A. B. Abdallah, and K. N. Dang, "An in-situ dynamic quantization with 3D stacking synaptic memory for power-aware neuromorphic architecture," IEEE Access, vol. 11, pp. 82377–82389, 2023.
[35] M. Evers, L. Barnes, and M. Clark, "The AMD next-generation 'Zen 3' core," IEEE Micro, vol. 42, no. 3, pp. 7–12, May 2022.
[36] K. Ueyoshi et al., "QUEST: Multi-purpose log-quantized DNN inference engine stacked on 96-MB 3-D SRAM using inductive coupling technology in 40-nm CMOS," IEEE J. Solid-State Circuits, vol. 54, no. 1, pp. 186–196, Jan. 2019.
[37] K. Shiba et al., "A 96-MB 3D-stacked SRAM using inductive coupling with 0.4-V transmitter, termination scheme and 12:1 SerDes in 40-nm CMOS," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 68, no. 2, pp. 692–703, Feb. 2021.
[38] C. Frenkel, C. Frenkel, M. Lefebvre, J.-D. Legat, and D. Bol, "A 0.086-mm2 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28nm CMOS," IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 1, pp. 145–158, Feb. 2019.
[39] T. Wunderlich et al., "Demonstrating advantages of neuromorphic computation: A pilot study," Frontiers Neurosci., vol. 13, p. 260, Mar. 2019.
[40] R. V. W. Putra, M. A. Hanif, and M. Shafique, "EnforceSNN: Enabling resilient and energy-efficient spiking neural network inference considering approximate DRAMs for embedded

systems," Frontiers Neurosci., vol. 16, Aug. 2022, Art. no. 937782
[41] N. Inc. Nangate Open Cell Library 45 nm. Accessed: Feb. 14, 2023.[Online]. Available: http://www.nangate.com/
[42] N. E. D. Automation. FreePDK3D45 3D-IC Process Design Kit.Accessed: Feb. 14, 2023. [Online]. Available: http://www.eda.ncsu.edu/wiki/FreePDK3D45/
[43] M. R. Guthaus, J. E. Stine, S. Ataei, B. Chen, B. Wu, and M. Sarwar,"OpenRAM: An open-source memory compiler," in Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD), Nov. 2016, pp. 1–6.
[44] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Can. Inst. Adv. Res., Toronto, On, Canada, Apr. 2009. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html
[45] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, "Cacti 6.0: A tool to model large caches," HP Laboratories, vol. 27, p. 28, Apr. 2009.
[46] P. Reviriego, P. Junsangsri, S. Liu, and F. Lombardi, "Error-tolerant data sketches using approximate nanoscale memories and voltage scaling," IEEE Trans. Nanotechnol., vol. 21, pp. 16–22, 2022.
[47] P. Royer and M. López-Vallejo, "Using pMOS pass-gates to boost SRAM performance by exploiting strain effects in sub-20-nm FinFET technologies," IEEE Trans. Nanotechnol., vol. 13, no. 6, pp. 1226–1233, Nov. 2014.
[48] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," IEEE J. Solid-State Circuits, vol. SSC-22, no. 5, pp. 748–754, Oct. 1987.
[49] M. Karimi, A. S. Monir, R. Mohammadrezaee, and B. Vaisband, "CTT-based scalable neuromorphic architecture," IEEE J. Emerg. Sel. Topics Circuits Syst., vol. 13, no. 1, pp. 96–107, Mar. 2023.
[50] A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm," Integration, vol. 58, pp. 74–81, Jun. 2017.