IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 July 2023, accepted 1 August 2023, date of publication 3 August 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301560

RESEARCH ARTICLE

An In-Situ Dynamic Quantization With 3D Stacking Synaptic Memory for Power-Aware Neuromorphic Architecture

NGO-DOANH NGUYEN[1], (Member, IEEE), XUAN-TU TRAN[2], (Senior Member, IEEE), ABDERAZEK BEN ABDALLAH[1], (Senior Member, IEEE), AND KHANH N. DANG[1], (Member, IEEE)

[1]Graduate School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu–Fukushima 965-8580, Japan
[2]VNU Information Technology Institute, Vietnam National University, Hanoi, Hanoi 10000, Vietnam

Corresponding author: Ngo-Doanh Nguyen (m5262108@u-aizu.ac.jp)

This work was supported in part by The University of Aizu Competitive Research funding under Grant 2023-P26; and in part by the Very Large-Scale Integration (VLSI) Design and Education Center, The University of Tokyo, Japan, in Collaboration with Synopsys Inc., and Cadence Design Systems Inc.

ABSTRACT Spiking Neural Networks (SNNs) show their potential for lightweight low-power inferences because they mimic the functionality of the biological brain. However, one of the major challenges of SNNs like other neural networks is memory-wall and power-wall when accessing data (synaptic weights) from memory. It limits the potential of spiking neural networks implemented on edge devices. In this paper, we present a novel spiking computing hardware architecture named NASH-3DM using 3D-IC-based stacking memory with power supply awareness to effectively decrease power consumption for AI-enabled edge devices. Instead of storing one or multiple weights in a single memory word, we split them into small subsets and allocate each subset into a separate memory in every stacking layer. With the natural separation of stack layers, our system can activate and deactivate each layer separately. Therefore, it can offer *in-situ* (online, post-manufacture, and without interruption) dynamic quantization with multiple operating modes. With the CMOS 45nm technology, our energy per synaptic operation for MNIST classification can reduce by 36.67% while having 0.93%-1.14% accuracy loss at 5-bit quantization. The energy per synaptic operation reduction for the CIFAR10 dataset is 36.68% when switching from the 16-bit active operation to the *in-situ* 10-bit one with an accuracy loss of 5.69%.

INDEX TERMS Spiking neural network, 3D IC-based stacking memory, digital neuromorphic.

I. 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], [3], [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

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

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 word 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, in this paper, we present a novel *in situ* dynamic quantization hardware architecture of a spiking computing processor using 3D-IC-based stacking memory. In our previous publications [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 paper are summarized in the following:

- A novel 3D-IC stacking synaptic memory architecture, called NASH-3DM, supports splitting the memory word into subsets. The SNN architecture supports computing with not only all subsets available but also missing subsets.
- An *in situ* dynamic quantization approach. In contrast to the conventional *ex-situ* quantization, the bit-width of synaptic weight is decided before fabrication and stays unchanged during inference. Our architecture supports changing the bit-width *in situ* and dynamically by turning on and off the memory layers. As a result, the system has multiple power modes and adjusts the bit precision of synaptic weights based on its power supply.
- A novel yield improvement approach for the proposed NASH-3DM by swapping the subsets once defects are detected in the layer. This approach acts like quantization by cutting off the least significant bit subset. Moreover, it can be a fail-safe feature in our system against new defects if necessary.
- Evaluating the transformation of power consumption, and accuracy of the 3D spiking computing processor at multiple bit-width modes. This evaluation is based on the NANGATE 45nm Open Cell Library [12] as the standard cells, OpenRAM library [13] for generating the system memory, and the Through-Silicon Via (TSV) from FreePDK3D45 [14] for 3D implementation.

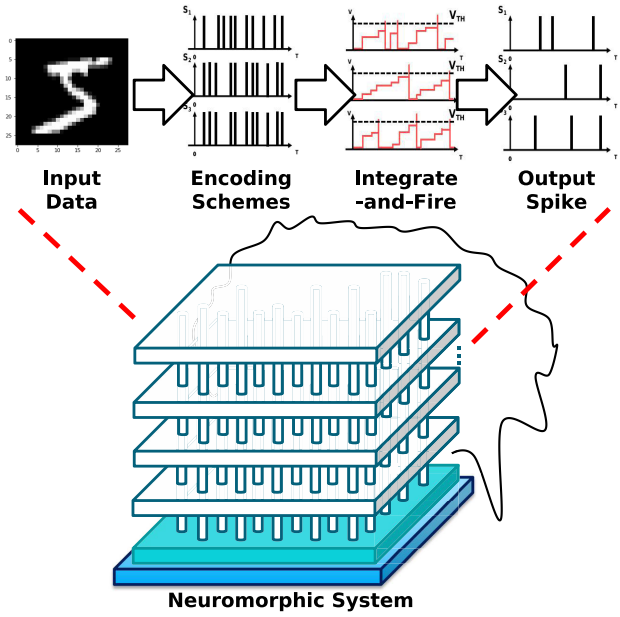

FIGURE 1. High-level view of the 3D-IC-based spiking neural network architecture.

The rest of this paper is organized as follows. Section II presents our motivation for the *in-situ* 3D spiking computing processor and explains its hardware architecture. In Section IV, the performance and power consumption of our NASH-3DM will be evaluated. Finally, there are some conclusions and perspectives in Section V.

II. BACKGROUND AND MOTIVATIONS

A. BACKGROUND

The high-level view of 3D-IC-based SNN architecture is shown in Fig. 1. Compared to other neural network models, information is encoded in Spiking Neural Networks (SNNs) using an encoding scheme. This information is then transmitted between neurons through trains of action potentials called spikes. Those spikes biologically are generated by the neuron's membrane potential reaching a certain threshold. They operate in a discrete-time domain, with each neuron sending and receiving spikes at specific times. As a result, it allows them to process temporal information, such as patterns and sequences, in a more natural way than traditional Artificial Neural Networks (ANNs). The most popular hardware model for simulating this behavior of biological neurons is the Leaky Integrated-and-Fire (LIF) because of its energy

efficiency and capability of capturing the essential features of bio-information. Theoretically, LIF neuron operations are expressed in the following equation:

$$V_i(t) = V_i(t-1) + \sum_j w_{i,j} \times x_j(t-1) - \lambda \quad (1)$$

where $w_{i,j}$ is the synaptic weight between the i^{th} neuron and the j^{th} one. $V_i(t)$ is the membrane potential of i^{th} neuron at the t timestep and $x_j(t-1)$ is the output pre-synaptic spike of i^{th} neuron and the leaky value λ, respectively. This output of the i^{th} neuron is expressed with the equation below.

$$x_i(t) = \begin{cases} 1, \text{ if } V_i(t) \geq V_{th}, \\ 0, \text{ otherwise.} \end{cases} \quad (2)$$

Moreover, the neuromorphic systems are expected to be asynchronous and independent of neurons within the network. Therefore, the ability to learn the timing information is also crucial. In practice, there are two learning approaches, which are off-chip learning and on-chip learning. For the off-chip method, the popular one is the ANN-to-SNN conversion with a fully connected feed-forward neural network using the RELU activation function [15]. It is usually trained in software using back-propagation with zero bias and then mapped into the LIF network in a normalized way. For the on-chip method, the famous algorithm is the Spike-Timing-Dependent Plasticity (STDP) [16], an unsupervised learning algorithm with the biological characteristic. It is based on synaptic plasticity to represent the relative difference in timing between the pre-synaptic spike and the post-synaptic one.

B. RELATED WORKS

1) ARCHITECTURE DESIGN PERSPECTIVE

In Section I, it was mentioned that there are three different approaches to designing SNN hardware. The most widely used approach is the *2D-ICs based digital hardware*. Notable examples of this approach include Intel's Loihi [2] and IBM's TrueNorth [5]. Loihi utilizes the asynchronous Network-on-Chip (NoC) to represent the spike transmission of active synapses. Furthermore, Loihi's neurons are reconfigurable, allowing for the implementation of different neuron models and supporting adaptive bit-width operations (1)-to-9 bits) for synapses. In the case of IBM's system, TrueNorth relies on fixed-bit-width weights for its Leaky-Integrated-and-Fire (LIF) neuron cores. However, TrueNorth operates on a large-scale network with 1 million neuron cores, each having a 256 × 256 crossbar connecting pre-synaptic spike events to post-synaptic ones. In conclusion, TrueNorth stands out due to its ease of prototyping and system debugging. However, in terms of power consumption, it requires more power than the other two approaches (*2D-ICs based analog mix-signal hardware* and *3D-ICs based hardware*) when scaled to the identical fabrication technology [17].

Regarding the *2D-ICs based analog mixed-signal hardware*, this approach has the ability to accurately emulate the electrical behaviors of biological neurons while having lower power consumption than digital systems. A demonstration of such a system is NeuroGrid from Stanford University [3], which is based on the analog sub-threshold design. This system is capable of achieving real-time performance. NeuroGrid utilizes the Network-on-Chip (NoC) with a tree topology and multicasting feature. Despite using older technology (180nm), NeuroGrid outperforms TrueNorth (28nm) in terms of energy efficiency, with an energy-per-operation result of 45pJ compared to 50pJ. Moreover, the analog mixed-signal approach can also match the capabilities of the digital system in cases of scalability and robustness, as demonstrated by Heidelberg University's BrainScaleS-2 architecture [4]. This system utilizes analog wafer-scale circuits and operates at a time scale 10.000× times faster than real-time biological processes. However, fabricating analog circuits have higher complexity than digital circuits. The reason is that standard analog cells tend to require customization when shifting technology. Additionally, these systems pose challenges in terms of control and calibration, even when scalability is achieved. This is due to significant variations in analog circuit characteristics across different process technologies, temperatures, and voltage levels.

In terms of the *3D-ICs based hardware*, there is growing interest in the Loihi-2 architecture [7], which supports 3D multi-chip scaling and represents the next generation of hardware architectures. NeuroSIM [6], a 3D neuromorphic system, incorporates two-layer memristors as electronic synapses for SNN. This integration leads to a 50% reduction in the hardware area, 1.48× times lower power consumption, and 2.58× times lower latency compared to traditional 2D single-layer configurations. Another 3D-IC-based SNN architecture called MigSpike [11] is specifically designed for fault tolerance and reduces migration costs associated with remapping in NoC by a factor of 10.19× compared to 2D approaches. Consequently, 3D-ICs offer significant advantages over the aforementioned approaches, including reduced hardware footprint, cost, and power consumption. It is reasonable to expect that a 3D SNN system would provide even greater benefits in terms of power consumption and hardware area reduction for edge devices.

2) POWER EFFICIENCY PERSPECTIVE

SNNs have attracted attention for their superior energy efficiency over other neural network models [18], [19]. At the moment, although other neural networks have been able to perform various cognitive tasks at the human level, the power consumption of those models deployed in the state-of-the-art hardware is still significantly higher than our human brain, which consumes only ∼20W for those tasks with a volume of ∼1200cm^3 [20]. Moreover, for real-time energy-hunger applications such as robotics, automobile, and Internet-of-Things (IoT), conventional deep neural networks (DNNs) with a large number of computations have become the main obstacle [21]. In contrast, neuromorphic systems emulate the ultra-high efficiency and agility of brains, along with their correctness [19]. The reason is that SNNs, which incorporate

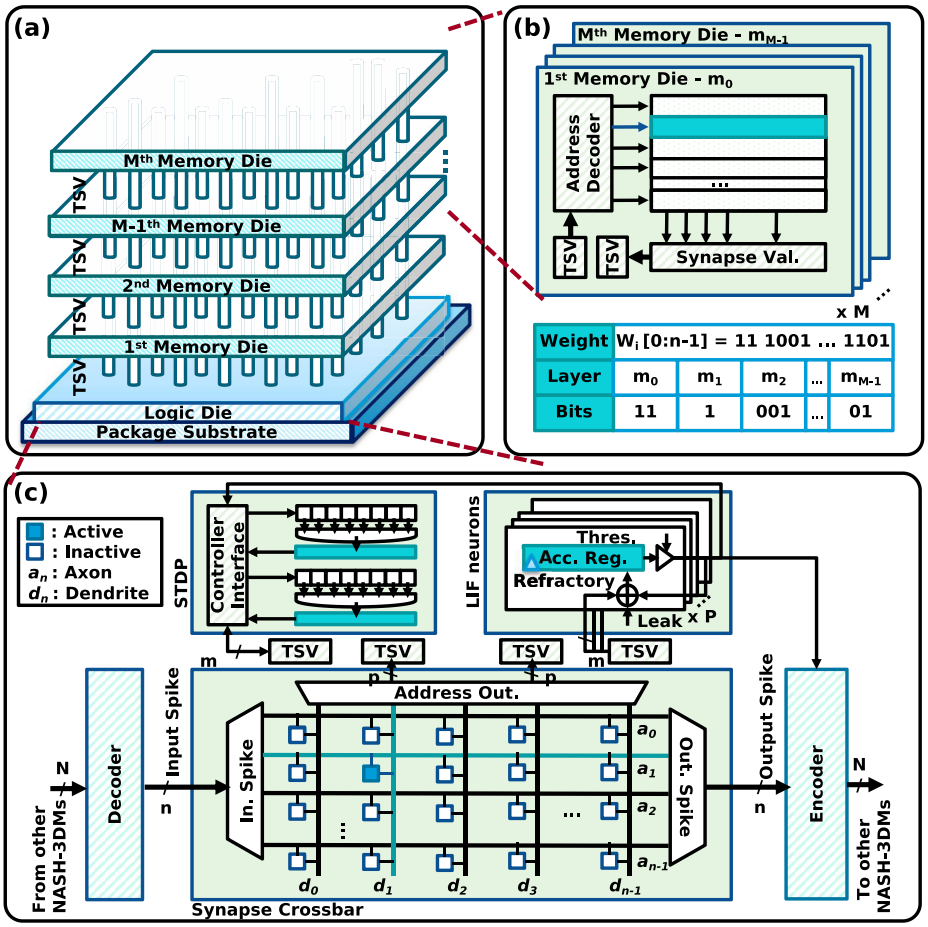

FIGURE 2. The overview hardware architecture of NASH-3DM with 3D-ICs based stacking memory. (a) The hardware contains *P* Leaky Integrate-and-Fire (LIF) cores and *M* memory layers stacked on top of it. (b) The bit distribution in *M* stacking memory layers. (c) The hardware architecture of each LIF core at the logic die.

temporal dynamics of spikes, only output when necessary and do not constantly process information like traditional neural networks. As a result, it is suitable for edge devices, which usually utilize the sensors to record temporal information in the environment [22]. As a result, SNNs become an emerging neural network model that could promisingly provide efficiency and reliability for AI-enabled edge devices.

An alternative approach to enhancing power efficiency involves utilizing new memory technologies, such as In-Memory Computing (IMC) and 3D stacking memory. IMC currently offers two approaches: analog IMC [23], [24], [25] and digital IMC [26], [27], [28]. Analog IMC, unfortunately, suffers from limited conversion accuracy due to low-cost analog-to-digital converters (ADC), making it unsuitable for applications demanding high precision, such as automobiles. Analog IMC also faces challenges related to variations caused by factors like temperature and sneak currents, which can affect its performance [29]. Conversely, digital IMC boasts higher computational accuracy and robustness but consumes more power compared to analog IMC [30]. On the other hand, 3D stacking memory aims to achieve greater memory capacity and minimize data movements [31], [32]. By stacking multiple SRAMs, substantial bandwidth and caching capacity can be attained for CPUs or DNN inferences [33], [34]. Communication between the stacked layers can be achieved through wired integration using through-silicon vias (TSVs) [31], [32], or wireless integration via inductive coupling, known as ThruChip Interface (TCI) [35], [36]. Regardless of the method chosen, 3D stacking memory holds the potential to enhance power efficiency by reducing data movements.

Combining the above point with the fixed-pattern-noise resilience of SNNs [9], we observe a chance to keep the synaptic operations at low-power mode, which is still able to maintain the performance of neuromorphic systems, to reduce overall power consumption by using the characteristic of 3D stacking memory and dynamic bit-width quantization. The next section will provide a detailed explanation of this 3D neuromorphic architecture.

III. HARDWARE ARCHITECTURE

Fig.2 illustrates the architectural overview of our NASH-3DM hardware. Here, we show the NASH-3DM contained *P* LIF neurons with *M* stacking memory layers. All neurons

or processing elements are placed at the bottom layer (logic die) and the stacked layers (memory die) contain only synaptic memory. The synaptic weights are partitioned into M memory layers, with data transmission via Through-Silicon Vias (TSVs). It is important to highlight that the number of LIF neurons and memory layers are customizable parameters that can be adjusted during the design phase. Each neuron inside NASH-3DM has its own address decoder and encoder inside to update the synaptic weights correctly. They act as the receiver and transmitter for messages in the network. The output spike of LIF neurons to the next ones could either be in the same NASH-3DM or other NASH-3DMs. On the contrary, the input spike received from the previous neurons triggers the crossbar to attach the corresponding weights from memory layers via TSV for the LIF function. Each LIF neuron contains one STDP for self-learning and self-updating synaptic weights over operating time.

A. 3D STACK MEMORY STRUCTURE WITH IN SITU DYNAMIC QUANTIZATION

Let's assume the SNN system uses n-bit weight format for design which stays unchanged after manufacturing. Rather than consolidating one or multiple n-bit weights within a single memory word, our approach involves dividing each p-bit weight into a collection of subset bits $\{m_0, m_1, \ldots m_{M-1}\}$, where m_i represents subset i and M denotes the total number of subsets. Notably, m_0 represents the subset with the highest significance, while m_{M-1} corresponds to the subset with the lowest significance.

The *in situ* dynamic quantization is obtained by following the rule:

- At the beginning and in normal power consumption mode, all sub-sets of synaptic weights are stored in all memories.
- If lower-power mode is enabled, the system starts to deactivate the LSB subset using the power-gating technique. In the processing elements, the turned-off subsets used in LIF computations are treated as zeros.
- If normal power mode is enabled, the system starts to turn on the subset containing the most significant bits among all inactive subsets.

In the exemplary model as in Fig. 2(b), we divide those $n = 8$-bit weights into M separated memory layers. The synaptic weights can be split unevenly into these layers. In addition, the LSBs are on the top memory layer(s) and the MSBs are on the bottom. By separating the bits of synaptic weights into different layers, our hardware architecture is capable of power-gating the top memory layer(s) to act as reducing the bit precision of SNN (called *in-situ* dynamic quantization). The LSBs will be treated as all zero in the processing elements. Consequently, this leads to a significant reduction in overall power consumption while maintaining a graceful level of accuracy. It is suitable for edge devices when their battery or power source almost runs out. This happens by taking advantage of the noise and bit-loss resilience of SNN, which other neural network models usually lose their accuracy sharply because of the operating-bit reduction. Moreover, with the separating structure, this approach has two other benefits. First, the quantization can be operated after manufacturing and without any interruptions in the system operations. Hence, in the case of the power supply reaching a certain low-level threshold, the system could switch to the low-power mode, which reduces a small fraction of accuracy, to increase the operation time. Second, unlike *ex-situ* quantization, the LSBs can be refilled and reattached if necessary during the operations. It is important because of the fact that the power supply can be also dynamically adjusted or recharged at run time.

B. SYNAPTIC WEIGHTS OPERATION WITH 3D STACKING MEMORY

As shown in Fig. 2(c), each NASH-3DM has its own synaptic crossbar, which is to enable the synapse weights to be read from the synapse memory and to update weights in parallel. For example, with N inputs of n bits, each LIF neuron utilizes a total of N $\times$ n bits in M SRAMs. As shown in Fig. 3, the subsets of synaptic weights could be either split equally or unequally. Although 1-bit words or 2-bit words can be unreasonable, we can pack several LIF neurons' synapses SRAM to have more traditional word sizes (8,16, 32, or more) [5], [37].

In each timestep, the crossbar checks for the appearance of input spike(s). If the spike appears, the one-hot mechanism decodes the address of corresponding synaptic weights located on $M = 4$ SRAMs via TSVs. TSVs here act as wires to connect between layers. The synaptic weights are then fetched from memory and loaded to the LIF neuron for the accumulating calculation. When the spike is absent, the accumulation is stopped, and the NASH-3DM moves to the next operation.

In each NASH-3DM, the synaptic crossbar can operate independently. However, the adapting operations are only available on the memory layers that still had their power supply. For the missing bits due to power-gating, these values are treated as zeros on the ongoing operations. Fig. 4 shows this operation in detail, in which the output spike event(s) may change according to the missing bits.

In Fig. 4, we provide an example where the synaptic weights have inactive bits because of power-gating memory layers. The LIF neurons accumulate the synaptic weights according to the input spike events. Whenever the accumulation reaches the pre-trained threshold, the LIF neuron fires the output spike to the downstream neurons. However, when the power supply is low, our *in-situ* NASH-3DM switches its operation mode by power-gating the memory layer one by one. The aim is to reduce the power consumption while taking fixed-pattern-noise-resilient advantages of SNNs to maintain accuracy [9]. In practice, the NASH-3DM applies the power-gate technique to memory layers with the top-down direction, which starts from the LSBs of synaptic weights and keeps the MSBs. For instance, by power-gating the top

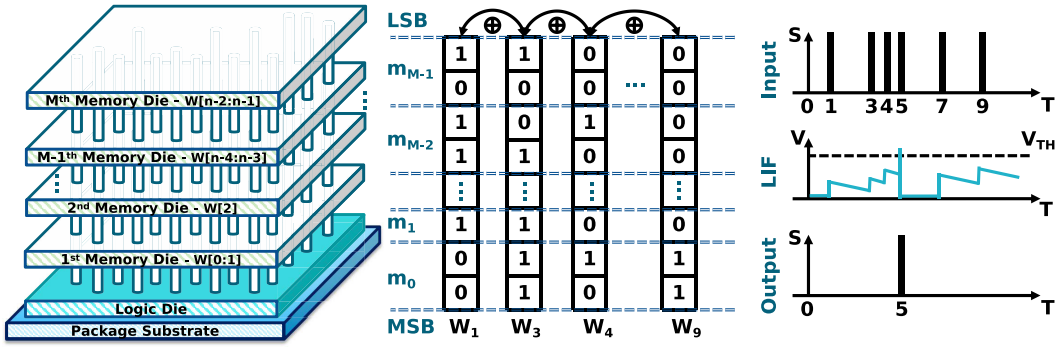

FIGURE 3. Demonstration of *n*-bit operations of synaptic weights under normal conditions (no power-gating).

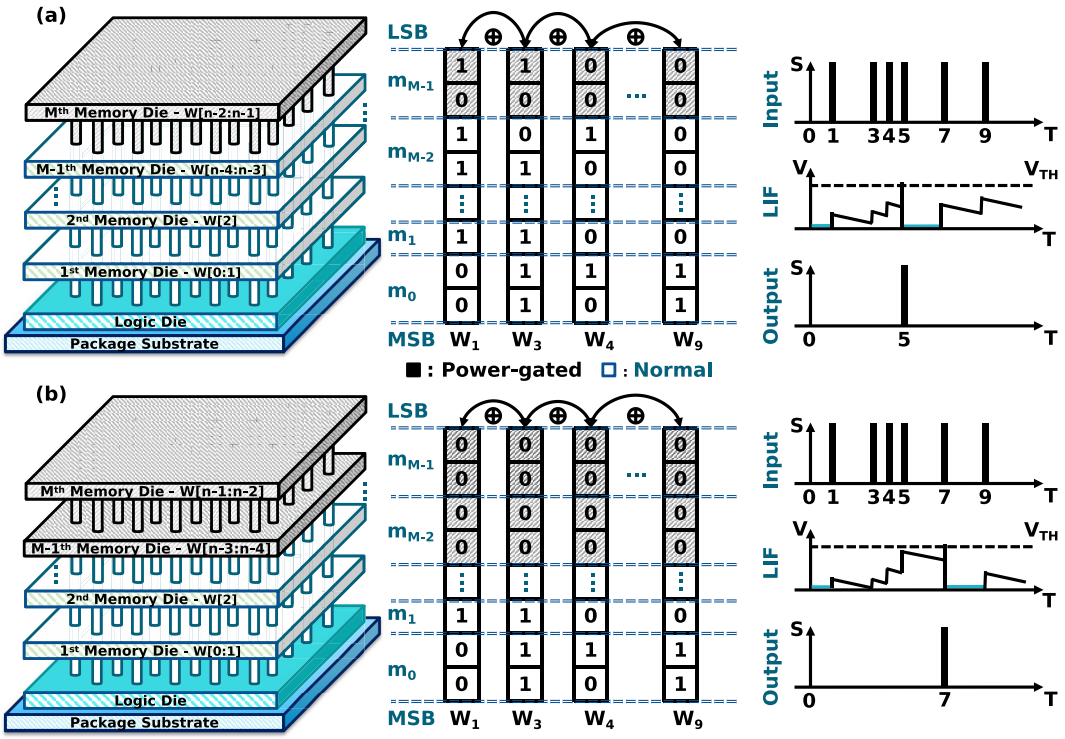

FIGURE 4. Demonstration of *n*-bit LIF operations of synaptic weights with *in-situ* quatization using power-gating. (a) The LIF operation of NASH-3DM without a power supply for the top memory layer. (b) The LIF operation of NASH-3DM without the power supply for two upper memory layers.

memory layer consisting of two LSBs, the LIF modules consequently operate on 6-bit synaptic weights.

In Fig. 4(b), two memory layers containing LSBs are discarded and the rest MSBs are still available to perform LIF operations. However, the output spike has been delayed because the accumulation cannot reach the threshold. The reasons are that the threshold stays the same as it is in the normal power mode and the four LSBs are now treated as zeros. Consequently, if the number of power-gated layers continuously increases, and the synaptic weights are removed at a certain level, the output spike event may not occur. Hence, it changes the sequence of the next computations in SNN in the wrong way, which affects the final prediction. Therefore, despite the noise resiliency of SNNs, there are still several cases that cause the alternation of the final results of NASH-3DM when the memory layers are power-gated. The change in weights could lead to the alternation of output spike events which may alternate the overall accuracy as a trade-off.

C. POWER CONSUMPTION FOR IN-SITU QUANTIZATION

In this subsection, we analyze the projection of the power efficiency of *in-situ* quantization for NASH-3DM with multiple stacked memory layers. Overall, there are two main power consumers in our NASH-3DM, which are P_{mem} from the memory and P_{pe} from the processing elements (LIF core, controller, STDP, address decoder, and encoder). It can be

expressed as the following equation.

$$P_{total} = P_{mem} + P_{pe} \tag{3}$$

However, a large part of power consumption in neuromorphic systems comes from memory, which is around 75% of total power [38]. Consequently, by altering the power distribution in memory, the total power consumption could alter significantly. Therefore, our method is to power-gate the memory layers for power reduction and it also provides an *in-situ* synaptic weight quantization. For example, with the 8-bit ($n = 8$) synaptic memory from the architecture in Fig. 2, we can define the total power consumption of synaptic memories with the following equations.

$$P_{mem} = P_{mem_{int}} + P_{mem_{leak}} + P_{mem_{switch}} \tag{4}$$

where $P_{mem_{int}}$ is the internal power of synaptic memories; $P_{mem_{leak}}$ is the leakage power of synaptic memories; $P_{mem_{switch}}$ is the power consumption of synaptic memories from switching activities. If we assume an equal distribution of power supply among synaptic memories, the power consumption of memory mathematically decreases by k/n when one or more memory layers, containing k LSBs, are power-gated.

$$P'_{mem} = \frac{n-k}{n} \times (P_{mem_{int}} + P_{mem_{leak}} + P_{mem_{switch}}) \tag{5}$$

This is due to the fact that all the memories in the layers are consolidated and exhibit identical switch activities when an input spike event occurs. With a value of $n = 8$ as shown in Fig.2, we can achieve anticipated power reductions of 25% and 50% for $k = 2$ and $k = 4$, respectively. Additionally, assuming that the memory accounts for approximately 80% of the power in a neuromorphic system, we could potentially reduce the overall power consumption by 20% and 40% correspondingly. As a result, for each feasible value of k, we can establish a power-aware mode.

Nonetheless, it is crucial to consider the remaining bits of synaptic weights to ensure the accuracy of the SNN model. Despite the noise resilience of SNNs, excessively power-gating the memory layers beyond a certain threshold will lead to the collapse of the spiking computing processor, rendering it unable to function correctly. In Section IV, the experimental results for each power-aware mode of operation will be presented to further demonstrate this.

D. YIELD IMPROVEMENT MECHANISM

As having low yield rates is a critical issue of 3D-ICs, in this section, we discuss the yield-rate improvement for 3D designs with our proposed architecture. The issue of low yield rate poses a significant challenge in 3D-stacking technology. Assuming the yield rate for a single layer (die) is $\gamma_{layer} < 1.0$, the yield rate of stacking D layers, denoted as γ_{D_layers}, can be expressed using the following equation.

$$\gamma_{D\_layers} = \prod_{i=0}^{D-1} \gamma_i \tag{6}$$

where D is the number of layers, γ_{D_layers} is the overall yield rate and γ_i is the yield rate of the i^{th} layer. Therefore, by stacking multiple layers on top of each other, the yield rate is much smaller than γ_{layer}, according to Eq. 6. To illustrate, let's consider a scenario where all layers possess the same yield rate, $\gamma_{layer} = 0.95$, and the number of stacked layers is $D = 3$. Consequently, the effective yield rate of the 3D-stacked chip is diminished to 0.85, which is 0.1 lower than that of a single-layer chip (0.95).

However, in our architecture, we split the synaptic weights into multiple memories and stack them on top of the logic layer. Therefore, if the defect in one memory layer affects the overall accuracy greatly, we could consider to power-gate that defective layer and swapping the bits of synaptic weights in that layer with the layer containing lower significant bits. Fig. 5 shows an example of an output spike affected by the fabrication-defective memory layers.

As shown in Fig. 5, the defective layer will cause the errors in logic functions of transistors, which are usually the stuck-bit or bridging faults. Therefore, the fabricated design cannot function properly. In the case of Fig.5(a), the output spike fires earlier than expected, which causes the incorrect logic function. The reason is that the $(M-2)^{th}$ layer has been defective in the manufacturing process. However, in our architecture, we can power-gate the $(M-2)^{th}$ layer that caused the logic faults for the output spike and swaps the bits in that layer to the $(M-1)^{th}$ layer. It is because the $(M-1)^{th}$ layer contains the less valued bits than the $(M-2)^{th}$ layer. Assuming that the $(M-1)^{th}$ layer does not affect the outcome of the system. As a result, the output spike fires correctly. Consequently, we could consider accepting the manufacturing faults to increase the yield rate while reducing a fraction of accuracy. It will be explained in Section IV.

In general, with D stacking memory layers, NASH-3DM has T defective memory layers caused by the alternation of output spikes and $D - T$ unharmful layers. Therefore, if we accept the defects in the manufacturing process, the actual yield rate is improved as shown in the following equation.

$$\gamma_{D\_layers} = \prod_{i=0}^{D-T} \gamma_i \tag{7}$$

For example, NASH-3DM has a total of $D = 4$ layers with 3 stacking memory layers and it has two defective memory layers that affect the output spikes. Normally, this product is considered as not working. However, with our architecture, we could accept and power-gate these defective layers. Consequently, the faulty values become zeros and they are treated as LSBs. The higher-valued bits are shifted to other unharmful layers. As a result, NASH-3DM still operates with acceptable accuracy and the yield rate is improved. By plugging in the numerical values into Eq.7, the resulting actual yield rate is approximately $Y_{actual} \approx 0.9025$, rather than 0.8145, thereby resulting in an enhanced overall yield rate.

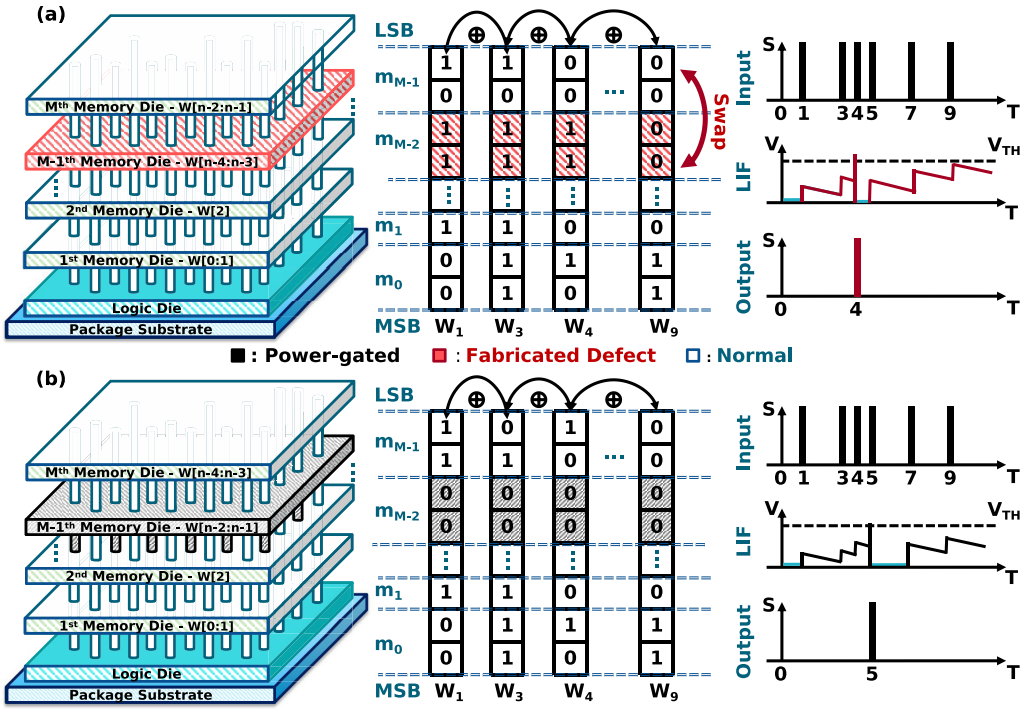

FIGURE 5. Example of Leaky-Integrate-and-Fire operations of NASH-3DM with defective memory layer(s). (a) The LIF of the NASH-3DM under one manufacturing-defected memory layer. (b) The LIF of NASH-3DM with power-gating the manufacturing-defected memory layer.

In summary, we have presented an approach to improve the low yield rate in 3D-IC-based SNN architecture. Furthermore, this mechanism can work as a fail-safe feature if new faults are detected which allows the system to maintain its operation under faulty situations.

IV. HARDWARE IMPLEMENTATION & EVALUATION

In this section, we perform the evaluation of the hardware complexity of NASH-3DM. Afterward, the quantization is evaluated in terms of power/energy and overall accuracy. Next, we evaluate the accuracy under defective memory layers before and after power-gating these layers. Finally, we compare our proposed architecture with the existing works. The proposed architecture was developed in Verilog, and the synthesis results and layout implementation were extracted using commercial Cadence tools. In addition, we use NANGATE CMOS 45nm open-cell library for ASIC implementation, OpenRAM for the system memory, and TSV from FreePDK3D45 for connecting between 3D layers.

To evaluate the power consumption, we use two image datasets, which are MNIST dataset [39] and the CIFAR10 dataset [40]. For the MNIST dataset, we use an 8-bit SNN structure consisting of 3 layers, which include one input layer, one hidden layer, and one output layer, and we change the size of the hidden layer. Hence, there are two settings for this structure, which are 784:512:10 and 784:64:10. For the CIFAR10 dataset, we use the VGG16 SNN model with 16-bit synaptic weights. The accuracy evaluation of the SNN models is based on off-chip learning. The images from the MNIST dataset and CIFAR10 dataset were transformed into spikes using rate coding with Poisson distribution. Please note that the configuration of this SNN structure can also be changed to different sizes and connections. In addition, the MNIST and CIFAR10 were selected for this evaluation because it has wide usage and provides a basic comparison with existing works [2], [5], [37], [41], [42], [43], [44].

TABLE 1. Hardware complexity of NASH-3DM.

Technology		45*nm*
Frequency		100*MHz*
# LIF		64 LIFs
# Stacking Memory		4 layers
# bit of Synaptic Weights		8 bits
Bit Configuration in Memory Layer		2-3-1-2
Gate Count	Total (NASH-3DM)	812.8*KGEs*
	Memory Blocks	791.76*KGEs*
	Crossbar & Address Decoder	9.68*KGEs*
	LIFs	11.36*KGEs*

A. HARDWARE COMPLEXITY

In this paper, we implement our proof-of-concept architecture with the NANGATE CMOS 45nm open-cell library with the support of OpenRAM for memory technology and FreePDK3D45 for TSV. Tab.1 shows the hardware area cost of our synthesized NASH-3DM for the MNIST dataset with 64 neurons in the hidden layer. It is about 812.8*KGEs* (kilo gate equivalents) at the operating frequency of 100*MHz*. Our

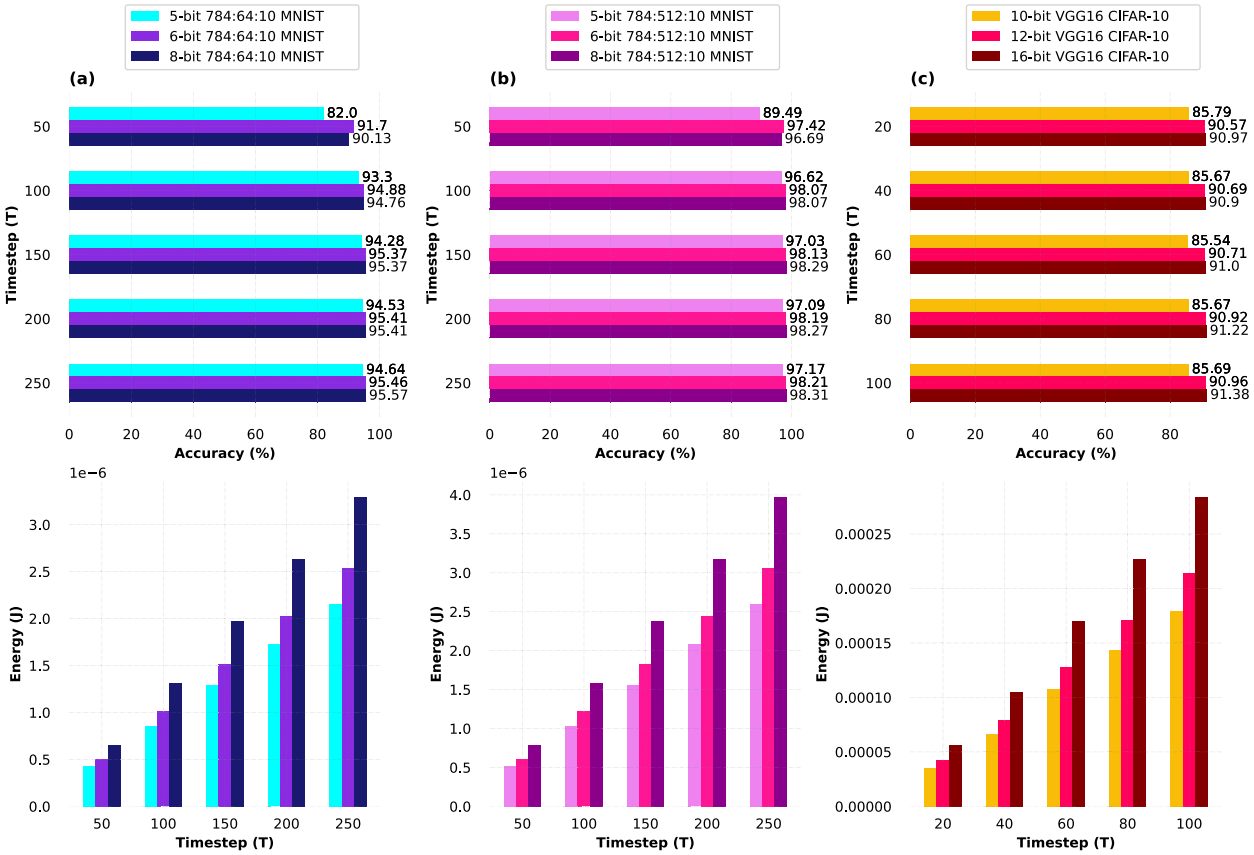

FIGURE 6. Energy Consumption and Accuracy of NASH-3DM for a prediction in different *in situ* dynamic quantization modes and different timesteps. (a) The evaluation is for the MNIST dataset with the layer configuration of 784:64:10. (b) The evaluation is for the MNIST dataset with the layer configuration of 784:512:10. (c) The evaluation is for the CIFAR10 dataset with VGG16 SNN model.

NASH-3DM uses 8 bits for synaptic weights. Moreover, we split those synaptic weights into four memory layers (m_0, m_1, m_2, m_3). The bit configurations of those layers are 2-bit, 3-bit, 1-bit, and 2-bit, respectively. The reason for this kind of division choice is that we want to keep the generalization of splitting weight bits by intentionally avoiding dividing the synaptic weight equally into 4 memory layers. We would like to note that the bit configurations are selected empirically in our experiment. Designers of course could choose different configurations (number of silicon layers, number of bits for each weight, and number of bits of each weight in each silicon layer); however, due to limited space and execution time, we choose the above configurations as a proof-of-concept work.

Specifically, the synaptic SRAM-based memory accounts for the majority of the hardware area, comprising approximately 97%. It is because we use the same size of SRAM in every memory layer even if the active bits in one layer are smaller than the actual size of SRAM. It is necessary to make space because it can use when switching defective layers, as mentioned in Section III-D. For the rest, the processing elements including crossbar, controller, and LIF neurons occupy 3% of the total hardware area of the NASH-3DM.

B. POWER VS. ACCURACY

As stated in Section III-B, our hardware area cost is unchanged with our *in situ* dynamical quantization approach. However, power consumption and accuracy are affected by switching the operation bits. Hence, we evaluate these transformations based on time steps. In detail, we analyzed our system's accuracy and energy under 8-bit, 6-bit, and 5-bit active operations for the MNIST dataset and 16-bit, 12-bit, and 10-bit active operations for the CIFAR-10 dataset on multiple time steps.

According to Tab.1, for the MNIST dataset, the 8-bit active operation is equal to the normal operation of NASH-3DM without power-gating any layers. Consequently, the power-gating top memory layer is the 6-bit active operation and the power-gating two top memory layers is the 5-bit active operation. In the case of the CIFAR-10 dataset, we keep the ratio of 16-bit synaptic operations as same as the 8-bit operations in the MNIST dataset. Therefore, by power-gating the top memory layer, the synaptic operations are based on 12 active bits; by power-gating the two upper memory layers, the synaptic operations are then based on 10 active bits. The evaluation of accuracy and energy consumption lasts from 50 to 250 timesteps for the MNIST dataset. For CIFAR10, the evaluation is from 20 to 100 timesteps.

As shown in Fig.6, for the MNIST dataset, the accuracy of our 8-bit-active NASH-3DM at the 250^{th} computing timestep reaches 95.57% and 98.21% with the SNN configuration of 784:64:10 and 784:512:10, respectively. With the 5-bit active operation, they drop by 0.93% and 1.14%, respectively. For the CIFAR10 dataset, the accuracy of 16-bit active operation is 91.38% at the 100^{th} computing timesteps and it drops by 5.69% when switching to 10-bit active operation. This highly suggests that there is a strong possibility for us to provide low-power modes while accepting a trade-off in terms of reduced accuracy.

In terms of energy, we evaluate the energy per prediction time-step-by-time-step, as shown in Fig. 6. For the total energy consumption with the same bit-width synaptic operation, the MNIST evaluation results increase from the 50^{th} timestep to the 250^{th} one by 5.041-5.055× fold. Similarly, the energy consumption with the VGG16 model increases approximately 5 times from the 20^{th} timestep to the 100^{th} one. On the other hand, those two energies with the MNIST dataset are dropped significantly by 22.95% and 34.44% at the 250^{th} timestep when we turn off one and two memory layers, respectively. For the CIFAR10 dataset, these numbers are 24.54% and 36.67% at the 100^{th} when we turn off one memory layer and two memory layers, respectively.

In summary, our evaluation results with MNIST demonstrate that our *in situ* dynamic quantization can reduce the energy per prediction by 36.67% while suffering 0.93%-1.14% of accuracy losses in 5-bit active operations with two SNN configurations (784:64:10 and 784:512:10). By combining with timesteps reduction, we can further reduce 87.04%, and 60.73% of energy reduction per prediction while suffering 7.2% and 1.18% for 5-bit active operation with SNN configuration of 784:512:10 in 50 and 150 timesteps, respectively. Our *in situ* dynamic quantization method has shown an excellent energy reduction ability.

C. YIELD RATE VS. ACCURACY

In this section, we evaluate the accuracy transformation under the defected probabilities according to three yield rates (0.905, 0.9905, and 0.99905). Assuming that the defective layer causes the stuck-bit event in the logic function of transistors and these defects have a uniform distribution. Therefore, the probability of the fault appearing in memory layers is equal to 0.095, 0.0095, and 0.00095, respectively. In this case, we use Monte Carlo simulation with 1,000 samples to get the average accuracy and its min-max values. The accuracy is evaluated with only the MNIST because it takes a long run time for CIFAR10 with the VGG16 SNN model. Hence, we use two SNN configurations for MNIST, which are 784:512:10 and 784:64:10.

As shown in Fig.7 and Fig.8, we illustrate the accuracy of NASH-3DM with three different fault probabilities, the normal accuracy and the accuracy with power-gating the defected memory layer. With the SNN configuration of 784:64:10, the worst-case accuracies for three yield rates (0.905, 0.9905, and 0.99905) at the 350^{th} are 95.25%, 95.37%, and 95.41%, respectively. Compared to the 8-bit operation without any defect, the accuracy reduces subsequently by 0.42%, 0.3%, and 0.26%. With the SNN configuration of 784:512:10, the worst-case accuracies change to 97.99%, 98%, and 98.04%, which reduce 0.32%, 0.31%, and 0.27% compared to the normal operation. However, when power-gating the defective memory layer, the accuracies in both SNN configurations (784:64:10 and 784:512:10) gain 0.21%-0.37% and 0.24%-0.29%, respectively. Therefore, we could accept the defective memory layer to improve the yield rate of 3D design while suffering a fraction of accuracy. Applying the numbers to Eq.6 and Eq.7, we improve the yield rate by about 0.0009-0.0638 with three assumptions (0.905, 0.9905, and 0.99905).

On the other hand, throughout all the computation time, we could observe that the accuracy starts to saturate around 100 timesteps. The impact on accuracy caused by defective synaptic weights in the early timesteps has a bigger impact than in the late timesteps. For example, as shown in both Fig.7 and Fig.8, at the timestep of 50, the original accuracy without any defects or noise is lower than other defective ones. The

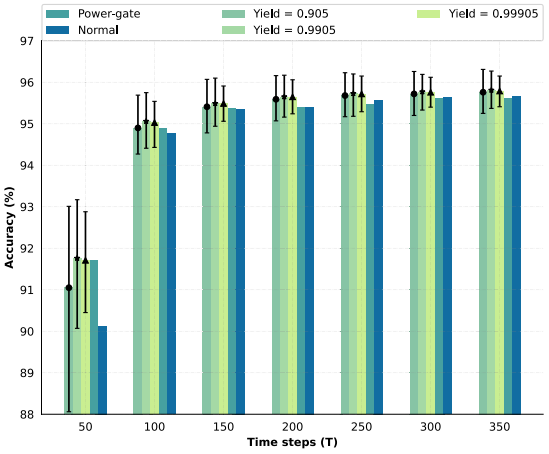

FIGURE 7. Accuracy of NASH-3DM (784:64:10) for MNIST dataset with different yield rates.

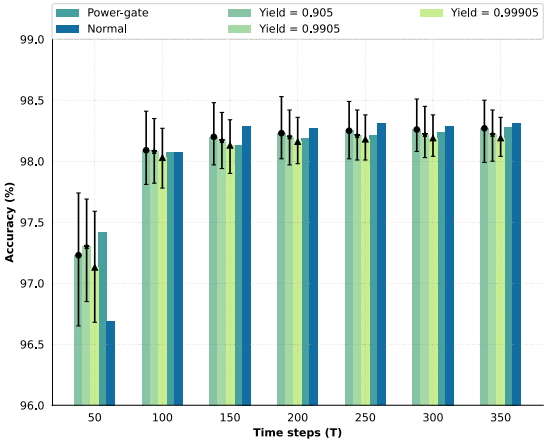

FIGURE 8. Accuracy of NASH-3DM (784:512:10) for MNIST dataset with different yield rates.

TABLE 2. Comparison results between the proposed architecture and existing works.

Parameters	TrueNorth [5]	Loihi [2]	ODIN [37]	NASH [10]	Karimi *et al.* [44]	NASH-3DM (this work)					
Benchmark	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST (784:64:10)			CIFAR10 (VGG16)		
Accuracy (%)	91.94	96	84	79.4	99.2	94.64	95.46	95.57	85.69	90.96	91.38
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		
In-situ quantization	N/A	N/A	N/A	N/A	N/A	5-bit	6-bit	8-bit	10-bit	12-bit	16-bit
Interconnect	2D	2D	2D	3D	2D	3D					
Implementation	Digital	Digital	Digital	Digital	Mix-signal	Digital					
Learning Rule	Un-supervised	On-chip STDP	On-chip Stochastic SDSP	On-chip STDP	Off-chip	On-chip STDP and 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	1.1V					
Energy per SOP (pJ)	26 (0.775V)	23.6 (0.75V)	8.4	189.3	8	160.01	188.04	244.08	300.96	358.58	475.20
Energy per SOP (pJ) (in 14nm)	4.902	23.6	1.078	10.86	4.32	9.18	10.79	14.01	17.27	20.58	27.27

reason is that we use rate coding in this evaluation and the number of timesteps affect the precision of the information transmitted between neurons. With more timesteps, the information becomes more detailed which leads to better accuracy and less noise impacts.

D. COMPARISON

In this section, we compare our hardware architecture with other existing works [2], [5], [10], [37], [44], as shown in Tab.2. We chose two SNN configurations which are fully-connected layers (784:64:10) for the MNIST dataset and VGG16 for the CIFAR10 dataset. For the MNIST dataset, we use the fixed 8-bit SRAMs for synaptic weights. For the CIFAR-10 dataset, we use the fixed 16-bit SRAMs for synaptic weights. Hence, we evaluate our architecture with three scenarios. They are the normal operations, the operations without the power supply for the top memory layer, and the operations without the power supply for the two upper memory layers.

Regarding accuracy, the evaluation reveals that our system achieves 95.57% in accuracy when applied to the MNIST dataset using an 8-bit operation. For the 5-bit operations, the accuracy drops by 0.93% compared to the 8-bit one. On the other hand, with the CIFAR10 dataset, our NASH-3DM achieves an accuracy of 85.68% and 91.38% with the 10-bit active operation and 16-bit active operation, respectively.

In relation to power consumption, we evaluate our work against others using the parameter of energy per synaptic operation. To account for the technology gap, we employ the scaling equation proposed by Stillmaker and Baas [45] to downscale the 14-nm technology node. The results, as presented in Tab.2, demonstrate that our hardware consumes 244.08pJ, 188.04pJ, and 160.01pJ for the MNIST dataset at the 45-nm technology node, utilizing 8-bit, 6-bit, and 5-bit active operations over 350 timesteps, respectively. When considering the CIFAR10 dataset, the energy per synaptic operation changes to 300.96pJ, 358.58pJ, and 475.20pJ for 10-bit, 12-bit, and 16-bit active operations. After scaling down to the 14-nm technology, our energy per synaptic operation achieves values of 9.18pJ, 10.79pJ, and 14.01pJ for the MNIST dataset accordingly. For CIFAR10 dataset, the numbers change to 17.27pJ, 20.58pJ, and 27.27pJ. Here, we could notice that CIFAR-10 implementation uses more energy per synaptic operation than MNIST because they utilized more bits in the synaptic weights (16 bits vs 8 bits) and in the *in-situ* dynamic quantization (10, 12, and 16 bits vs 5, 6 and 8 bits).

Compared to other works, our energy consumption has bigger values. It is because we design our memory layer to have the same size as SRAM to cover the defect from manufacturing, which causes extra power consumption. However, these evaluations indicate that our architecture, featuring 3D stacking memory, offers a notable benefit in terms of lowering energy consumption during the transition between operating modes.

V. CONCLUSION

In this paper, we have proposed a spiking computing processor with 3D-IC-based stacking synaptic memory named NASH-3DM and implemented it as a proof-of-concept system. The proposed architecture aims to the edge applications which their power supply tends to decrease over the operating time. With the 3D stacking memory, the neuromorphic

system can turn on/off the power supply of one or multiple metal layers inside it based on the power source. As a result, it can maintain a graceful degradation in accuracy while offering a low-power operation mode in power-hungry situations. In addition, our architecture could improve the yield rate by power-gating the defective memory layer(s). It is acted as a fail-safe feature of our hardware system against manufacturing defects. The defective layer(s) will be power-gated and the operating bits in that layer will be shifted to the upper memory layer(s). The reason is that the less important bits in the upper layers can be discarded and replaced by the more important bits in the lower layer(s). As a result, the defective hardware is able to continuously operate by reducing a small fraction of accuracy, which leads to yield improvement.

However, our proposed work still has some drawbacks. First, the combination of splitting weight bits is sub-optimal. It is because there are many combinations to divide synaptic weights. To address this problem, our future works will investigate the optimization algorithms such as Genetic Algorithms or Particle Swarm Optimization. Second, adding more low-power techniques (e.g., voltage-scaling, clock-gating) can further improve this work. Hence, one of our future works will be a combination of our quantization with lowering the memory voltage to further reduce the power consumption and the integration of large-scale systems using Network-on-Chips. Third, the 3D stacking SRAM used in this paper is only one of many memory types and our proposed architecture is able to work with any types of memory. In the future, we will investigate and implement other memory technologies such as 2D SRAM, ReRAM, and go on, to evaluate their power consumption in our system. Furthermore, we would like to investigate our yield improvement mechanism as a fail-safe future with the help of fault detection or testing.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful and valuable comments.

REFERENCES

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

[2] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, and G. Dimou, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018.

[3] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, May 2014.

[4] W. Guo, M. E. Fouda, A. M. Eltawil, and K. N. Salama, "Neural coding in spiking neural networks: A comparative study for robust neuromorphic systems," *Frontiers Neurosci.*, vol. 15, pp. 1–12, Mar. 2021.

[5] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[6] H. An, M. S. Al-Mamun, M. K. Orlowski, L. Liu, and Y. Yi, "Three-dimensional neuromorphic computing system with two-layer and low-variation memristive synapses," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 3, pp. 400–409, Mar. 2022.

[7] G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies, "Efficient neuromorphic signal processing with Loihi 2," in *Proc. IEEE Workshop Signal Process. Syst. (SiPS)*, Oct. 2021, pp. 254–259.

[8] A. Ben Abdallah and K. N. Dang, "Toward robust cognitive 3D brain-inspired cross-paradigm system," *Frontiers Neurosci.*, vol. 15, pp. 1–10, Jun. 2021.

[9] T. Wunderlich, A. F. Kungl, E. Müller, A. Hartel, Y. Stradmann, S. A. Aamir, A. Grübl, A. Heimbrecht, K. Schreiber, D. Stöckel, C. Pehle, S. Billaudelle, G. Kiene, C. Mauch, J. Schemmel, K. Meier, and M. A. Petrovici, "Demonstrating advantages of neuromorphic computation: A pilot study," *Frontiers Neurosci.*, vol. 13, pp. 1–13, Mar. 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. 64331–64345, 2021.

[11] K. N. Dang, N. A. V. Doan, and A. B. Abdallah, "MigSpike: A migration based algorithms and architecture for scalable robust neuromorphic systems," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 2, pp. 602–617, Apr. 2022.

[12] Nangate Inc. *Nangate Open Cell Library 45 nm*. Accessed: Feb. 14, 2023. [Online]. Available: http://www.nangate.com/

[13] 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. Computer-Aided Design (ICCAD)*, Nov. 2016, pp. 1–6.

[14] N. E. D. Automation. *FreePDK3D45 3D-IC Process Design Kit*. [Online]. Available: http://www.eda.ncsu.edu/wiki/FreePDK3D45

[15] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[16] M. R. Azghadi, N. Iannella, S. F. Al-Sarawi, G. Indiveri, and D. Abbott, "Spike-based synaptic plasticity in silicon: Design, implementation, application, and challenges," *Proc. IEEE*, vol. 102, no. 5, pp. 717–737, May 2014.

[17] S. Furber, "Large-scale neuromorphic computing systems," *J. Neural Eng.*, vol. 13, no. 5, Oct. 2016, Art. no. 051001.

[18] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[19] D. Wu, X. Yi, and X. Huang, "A little energy goes a long way: Build an energy-efficient, accurate spiking neural network from convolutional neural network," *Frontiers Neurosci.*, vol. 16, pp. 1–6, May 2022.

[20] C. D. James, "Toward exascale computing through neuromorphic approaches," Sandia, Livermore, CA, USA, Tech. Rep., SAND2010-6312, 2010.

[21] M. Pfeiffer and T. Pfeil, "Deep learning with spiking neurons: Opportunities and challenges," *Frontiers Neurosci.*, vol. 12, pp. 1–22, Oct. 2018.

[22] H. Mostafa, "Supervised learning based on temporal coding in spiking neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3227–3235, Jul. 2018.

[23] E. Lee, T. Han, D. Seo, G. Shin, J. Kim, S. Kim, S. Jeong, J. Rhe, J. Park, J. H. Ko, and Y. Lee, "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.

[24] M. E. Sinangil, B. Erbagci, R. Naous, K. Akarvardar, D. Sun, W.-S. Khwa, H.-J. Liao, Y. Wang, and J. Chang, "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.

[25] 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.

[26] 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.

[27] 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.
[28] 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.
[29] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams, "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Proc. 53rd ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2016, pp. 1–6.
[30] M. R. Haq 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.
[31] K. Cho, J. Park, B. Koo, S. Seo, Y. Hwang, S. Park, and M. Noh, "SAINT-S: 3D SRAM stacking solution based on 7 nm TSV technology," in *Proc. IEEE Hot Chips Symp.*, Mar. 2020, pp. 1–13.
[32] S.-K. Seo, C. Jo, M. Choi, T. Kim, and H.-E. Kim, "CoW package solution for improving thermal characteristic of TSV-SiP for AI-inference," in *Proc. IEEE 71st Electron. Compon. Technol. Conf. (ECTC)*, Jun. 2021, pp. 1115–1118.
[33] M. Evers, L. Barnes, and M. Clark, "The AMD next-generation 'Zen 3' core," *IEEE Micro*, vol. 42, no. 3, pp. 7–12, Feb. 2022.
[34] K. Ueyoshi, K. Ando, K. Hirose, S. Takamaeda-Yamazaki, M. Hamada, T. Kuroda, and M. Motomura, "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.
[35] K. Shiba, T. Omori, K. Ueyoshi, S. Takamaeda-Yamazaki, M. Motomura, M. Hamada, and T. Kuroda, "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.
[36] K. Shiba, M. Okada, A. Kosuge, M. Hamada, and T. Kuroda, "A 7-nm FinFET 1.2-TB/s/mm^2 3D-stacked SRAM Module With 0.7-pJ/b inductive coupling interface using over-SRAM coil and manchester-encoded synchronous transceiver," *IEEE J. Solid-State Circuits*, vol. 58, no. 7, pp. 2075–2086, Jul. 2023.
[37] C. Frenkel, M. Lefebvre, J. D. Legat, and D. Bol, "A 0.086-mm^2 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28 nm CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 1, pp. 145–158, Nov. 2018.
[38] 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, pp. 1–22, Aug. 2022.
[39] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.
[40] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Canadian Inst. Adv. Res., Toronto, Apr. 2009. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html
[41] J.-S. Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, B. Rajendran, J. A. Tierno, L. Chang, D. S. Modha, and D. J. Friedman, "A 45 nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2011, pp. 1–4.
[42] J. K. Kim, P. Knag, T. Chen, and Z. Zhang, "A 640 M pixel/s 3.65 mW sparse event-driven neuromorphic object recognition processor with on-chip learning," in *Proc. Symp. VLSI Circuits (VLSI Circuits)*, Jun. 2015, pp. C50–C51.
[43] J. Schemmel, D. Briiderle, A. Griibl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 1947–1950.
[44] 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.
[45] 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.

NGO-DOANH NGUYEN (Member, IEEE) is currently pursuing the master's degree with the Graduate School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan. He was with Vietnam National University, Hanoi, from 2018 to 2022, as a Research Engineer of system integration and VLSI design for artificial intelligent. His research interests include hardware/software co-design and verification and low-power solutions for artificial intelligence.

XUAN-TU TRAN (Senior Member, IEEE) received the Ph.D. degree in micro nano electronics from Grenoble INP, CEA-LETI, France, in 2008.

He was an invited Professor with the University of Paris-Sud 11, France, in 2009, 2010, and 2015; Grenoble INP, France, in 2011; and The University of Electro-Communications, Japan, in 2019. He was an Adjunct Professor with UTS, Australia, from 2017 to 2020. He is currently an Associate Professor with Vietnam National University, Hanoi (VNU). He is also the Director of the VNU Information Technology Institute (VNU-ITI). His research interests include design and test of systems-on-chips, networks-on-chips, design-for-testability, asynchronous/synchronous VLSI design, low power techniques, and hardware architectures for multimedia applications, cryptography. He is a Senior Member of the IEEE Circuits and Systems (CAS) and the IEEE Solid-State Circuits and Systems (SSCS). He is a member of IEICE and the Executive Board of the Radio Electronics Association of Vietnam (REV). He serves as the Chairperson for the IEICE Vietnam Section and the IEEE SSCS Vietnam Chapter.

ABDERAZEK BEN ABDALLAH (Senior Member, IEEE) received the Ph.D. degree in computer engineering from The University of Electro-Communications, Tokyo, in 2002. From April 2014 to March 2022, he was the Head of the Computer Engineering Division, The University of Aizu, Japan, where he has been the Dean of the School of Computer Science and Engineering, since April 2022, and a Full Professor. He is the author of four books, four registered and eight provisional Japanese patents, and more than 150 publications in peer-reviewed journal articles and conference papers. His research interests include adaptive/self-organizing systems, brain-inspired computing, interconnection networks, and AI-powered cyber-physical systems. He is a Senior Member of ACM.

KHANH N. DANG (Member, IEEE) received the M.Sc. degree from the University of Paris XI and the Ph.D. degree from The University of Aizu.

He is currently an Associate Professor with the Department of Computer Science and Engineering, The University of Aizu. His research interests include network-on-chips, 3D-ICs, neuromorphic computing, and fault-tolerant systems.