

# OSA-HCIM: On-The-Fly Saliency-Aware Hybrid SRAM CIM with Dynamic Precision Configuration

Yung-Chin Chen<sup>\*†</sup>, Shimpei Ando<sup>\*</sup>, Daichi Fujiki<sup>\*</sup>, Shinya Takamaeda-Yamazaki<sup>‡</sup>, Kentaro Yoshioka<sup>\*</sup>

<sup>\*</sup>Keio University, Japan, <sup>†</sup>National Taiwan University, Taiwan, <sup>‡</sup>The University of Tokyo, Japan

jim.chen.work@gmail.com, {shimpeiando, dfujiki}@keio.jp, shinya@is.s.u-tokyo.ac.jp, kyoshioka47@keio.jp

**Abstract**—Computing-in-Memory (CIM) has shown great potential for enhancing efficiency and performance for deep neural networks (DNNs). However, the lack of flexibility in CIM leads to an unnecessary expenditure of computational resources on less critical operations, and a diminished Signal-to-Noise Ratio (SNR) when handling more complex tasks, significantly hindering the overall performance. Hence, we focus on the integration of CIM with Saliency-Aware Computing—a paradigm that dynamically tailors computing precision based on the importance of each input. We propose On-the-fly Saliency-Aware Hybrid CIM (OSA-HCIM) offering three primary contributions: (1) On-the-fly Saliency-Aware (OSA) precision configuration scheme, which dynamically sets the precision of each multiply-and-accumulate (MAC) operation based on its saliency, (2) Hybrid CIM Array (HCIMA), which enables simultaneous operation of digital-domain CIM (DCIM) and analog-domain CIM (ACIM) via split-port 6T SRAM, and (3) an integrated framework combining OSA and HCIMA to fulfill diverse accuracy and power demands.

Implemented on a 65nm CMOS process, OSA-HCIM demonstrates an exceptional balance between accuracy and resource utilization. Notably, it is the first CIM design to incorporate a dynamic digital-to-analog boundary, providing unprecedented flexibility for saliency-aware computing. OSA-HCIM achieves a 1.95x enhancement in energy efficiency, while maintaining minimal accuracy loss compared to DCIM when tested on CIFAR100 dataset.

**Index Terms**—Computing-in-Memory (CIM), saliency, hybrid CIM (HCIM)

## I. INTRODUCTION

The massive data communication between memory and computing units presents a significant challenge to the efficient hardware acceleration of Deep Neural Networks (DNNs). This often results in increased energy consumption and protracted processing times. A promising solution to this issue is Computing-in-Memory (CIM), a technology that integrates computational functions into the memory array, thereby reducing data movement and significantly improving energy efficiency. Though general-purpose CIM technology has matured with extensive research and macro-level implementations [1], there is an ongoing shift towards developing CIM solutions that exploit network characteristics to achieve further performance breakthroughs. Such endeavors include leveraging DNN data distributions [2] and designing for specific network architectures [3].

This paper explores *Saliency-Aware Computing*, a software-driven mixed-precision computation paradigm which dynamically tailors computing precision based on the saliency, or the importance, of each input pixel. Figure 1 shows the application

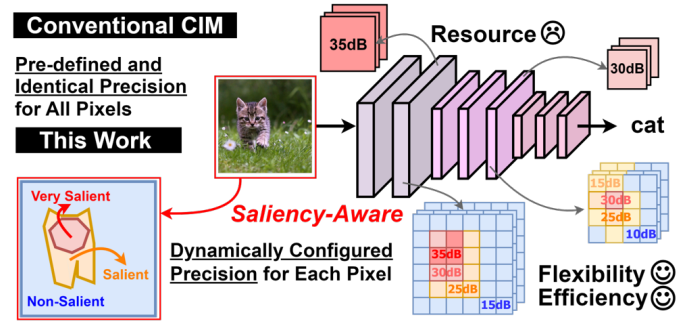


Fig. 1. Motivation of Saliency-Aware Computing. Each input pixel has a distinct impact on the output result of the DNN, necessitating different precision configurations. By dynamically employing high signal-to-noise ratio (SNR) precision for salient inputs and low SNR precision for non-salient inputs, efficiency can be enhanced.

of Saliency-Aware Computing to an image recognition task, using an image of a cat as an example. It is crucial to note that not all pixels hold equal importance; for instance, the pixels composing the cat's face and body are vital for recognition, while the background pixels depicting grass and flowers are largely irrelevant. By *dynamically* adjusting the computation precision in line with input saliency, computational resources can be effectively utilized. Specifically, we focus on high-precision computation for salient inputs that greatly impact classification accuracy, and lower-precision computation for non-salient inputs that contribute less. This selective allocation of computational resources allows for significant reductions in computational costs while maintaining high accuracy levels.

Despite the potential of saliency-aware computing to significantly enhance CIM performance, conventional CIMs lack the functionality to enable dynamic computing precision configuration. One contributing factor is the limitations within the vector-accumulation circuitry. For example, Digital-domain CIM (DCIM) necessitates excessive computation resources [4], [5], and Analog-domain CIM (ACIM) often compromises accuracy [6], [7]. Recent studies have attempted to integrate the digital and analog circuitries within the same macro unit [8], [9]. However, these hybrid schemes rely on fixed digital-to-analog configurations, which poses limitations on flexibility. To fully leverage the benefits of saliency-aware computing, a CIM design should have the capacity to 1) configure its multiply-and-accumulate (MAC) precision real-time with the saliency levels of input activation, and 2) allow dynamic

configuration of the digital-to-analog computation ratios.

To tackle these challenges, we introduce OSA-HCIM: a software-hardware co-designed hybrid CIM architecture. Differing from previous works, our approach synergistically combines the merits of software and hardware strategies. We specifically leverage the largely untapped potential of saliency-aware computing in the realm of CIM, and address the current limitations in dynamically configuring precision in hybrid CIMs. OSA-HCIM is characterized by three key features: (1) **Software Realm:** We introduce an On-the-fly Saliency-Aware (OSA) precision configuration scheme, providing numerous precision configurations for the MAC operation based on the online-evaluated saliency. (2) **Hardware Realm:** We propose a Hybrid CIM Array (HCIMA) capable of performing both bit-serial DCIM and bit-parallel ACIM concurrently using split-port 6T SRAM. (3) **Software-Hardware Co-design:** We present a comprehensive framework that integrates OSA scheme into HCIMA using a near-memory On-the-fly Saliency Evaluator (OSE) to effectively address various accuracy and efficiency requirements. The  $64 \times 144$  6T SRAM OSA-HCIM macro, implemented using 65 nm CMOS technology, stands out as the first CIM work to harness input saliency and incorporate a dynamic digital-to-analog computing boundary. The result demonstrates a 1.95x enhancement in energy efficiency while preserving similar levels of accuracy when compared to the full-digital approach on the CIFAR100 dataset.

## II. PRELIMINARIES

### A. Saliency-Aware Compute

The concept of saliency, which represents the varying importance of inputs, has been extensively utilized in various computer vision tasks such as video compression [10], super resolution [11], and object recognition [12]. This feature can also be leveraged in DNN tasks, as the contribution of each input activation to the final output result can vary significantly. For instance in Fig. 1, pixels containing the object of interest are more salient than the background pixels. To address this, saliency-aware computation have been developed. The computational precision is adjusted based on the input's saliency, where higher precision to salient pixels and lower precision to less significant ones is assigned. This strategy can reduce computational costs while minimally impacting the accuracy results.

Precision Gating (PG) [13] is a dual-precision software technique that assesses saliency by performing MAC operations between high-order input bits and weights, while DRQ [14] is a hardware architecture that calculates saliency by employing a mean filter on the input. However, these approaches have limitations: (1) support only dual precision configurations, providing limited tradeoff efficacy, and (2) lack memory-centric hardware solution.

### B. Digital, Analog, and Hybrid CIM

CIM architectures incorporate computing circuits within memory cells to perform MAC operations. Digital-domain CIM (DCIM) utilizes bulky digital adder trees (DATs) to

perform loss-free accumulations [4], [5], while analog-domain CIM (ACIM) employs analog accumulation technique, such as charge-sharing, along with an ADC conversion to achieve efficient but less accurate accumulation [6], [7]. To balance accuracy and resource consumption, recent research has explored the use of hybrid accumulation strategies that adopt both digital and analog schemes. This approach bifurcates the multi-bit MAC operation into DCIM and ACIM based on the input or output order using a pre-set threshold. Existing works either partition MAC operations into separated analog and digital CIM blocks based on pre-defined constraints [15], or implement both DCIM and ACIM within a macro with hard-wired boundary between digital and analog MAC computation [8], [9].

Nonetheless, the boundaries between digital and analog computations in these works are predetermined, and therefore lack adaptability to the varying precision requirements of each input. As a result, accuracy suffers due to analog noise when the input is more salient, while excessive resources are allocated for trivial inputs, degrading efficiency. To overcome this challenge, there is a need for a dynamically-configurable hybrid CIM capable of adjusting the digital and analog ratio according to the nature of the input patterns.

## III. SOFTWARE REALM: ON-THE-FLY SALIENCY-AWARE (OSA) PRECISION CONFIGURATION SCHEME

We propose the On-the-fly Saliency-Aware (OSA) precision configuration scheme, which introduces a dynamic computing precision mechanism optimized for CIM architectures. By evaluating the saliency of each input in real-time, OSA dynamically adjusts the ratio of digital and analog computation, thus enabling efficient resource utilization.

The OSA scheme operates within a standard CIM system that is optimized for binary (or 1-bit) MAC operations. To facilitate multi-bit MAC operations in such CIMs,  $a$ -bit input activations  $\vec{A}[a-1:0]$  and  $w$ -bit weights  $\vec{W}[w-1:0]$  are decomposed into  $w \times a$  1-bit MACs with output order  $k = i+j$  between 1-bit weights  $\vec{W}[i]$  and 1-bit activation  $\vec{A}[j]$ , where  $i$  ( $0 \leq i \leq w-1$ ) and  $j$  ( $0 \leq j \leq a-1$ ) indicate the bit order of weight and activation, respectively [16]. This decomposition is described in the equation below:

$$MAC(\vec{A}, \vec{W}) = \sum_{i=0}^{a-1} \sum_{j=0}^{w-1} 2^{i+j} \cdot MAC(\vec{A}[i], \vec{W}[j]) \quad (1)$$

Figure 2(a) illustrates the proposed OSA precision configuration scheme integrated into the decomposed multi-bit Hybrid CIM MAC operation. As an example, we consider a case where  $w = a = 6$ , with 36 distinct 1-bit MACs in total. To evaluate the saliency of the entire multi-bit MAC operation, the first  $s$  ( $s=2$  in Fig. 2(a)) highest-order 1-bit MACs, corresponding to MACs with output order  $k = w+a-2 \sim w+a-1-s$ , are computed precisely using DCIM. Subsequently, the output saliency of the multi-bit MAC is estimated based on these high-order 1-bit MAC results, which in turn determines the digital-to-analog boundary,  $B_{D/A}$ . Consequently, each 1-bit

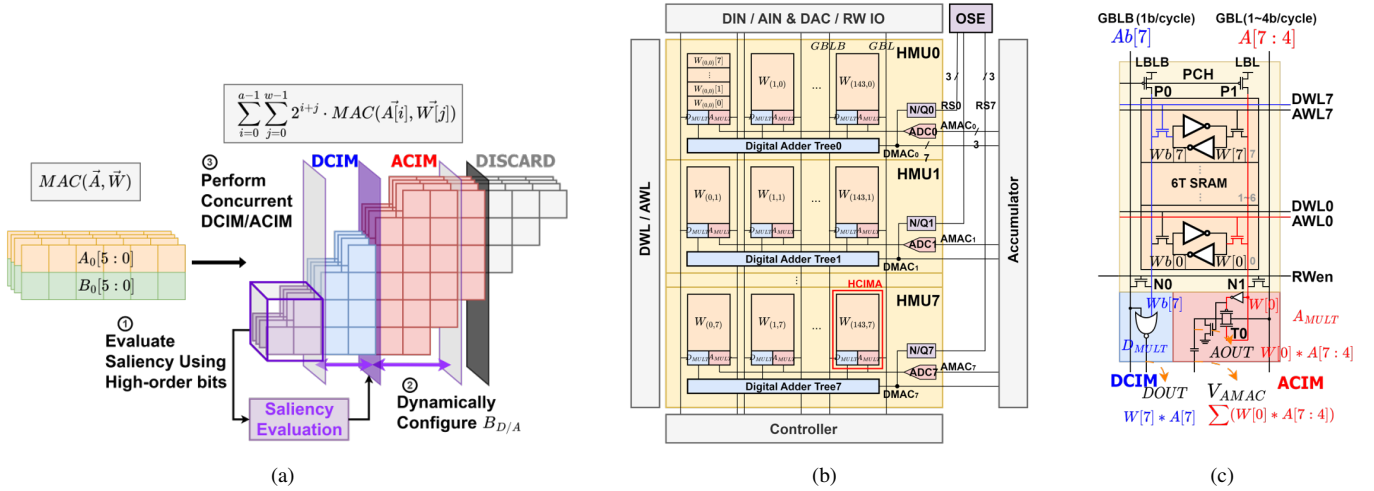


Fig. 2. (a) On-the-fly Saliency-Aware (OSA) Precision Configuration Scheme.  $B_{D/A}$  is dynamically configurable, allowing for a tradeoff between accuracy and efficiency based on input saliency. (b) The overall OSA-HCIM architecture. (c) Schematic of HCIMA.

MAC is categorized into *digital mode*, *analog mode*, or *discard* based on its output order  $k$ . MACs with  $k \geq B_{D/A}$  are computed with no accuracy compromise using DCIM, while those with  $B_{D/A} - 4 \leq k < B_{D/A}$  are computed with energy-efficient ACIM, albeit with noise incorporation. To further save power, MACs with  $k < B_{D/A} - 4$  are discarded due to their negligible impact on the final output. By dynamically configuring  $B_{D/A}$ , a unique precision configuration can be achieved, which enables the exploration of the optimal accuracy-efficiency tradeoff.

#### IV. HARDWARE REALM: OSA-HCIM ARCHITECTURE

##### A. OSA-HCIM Architecture Overview

The proposed OSA-HCIM hardware architecture features concurrent CIM operation across both the digital and analog domains. OSA-HCIM supports parallel execution within these hybrid domains, while providing the requisite flexibility for implementing our OSA precision configuration scheme. The culmination of these features leads to a substantial enhancement in computational performance and efficiency.

Figure 2(b) provides an architecture overview of OSA-HCIM. It is composed of the OSA-HCIM macro and its peripherals, including an On-the-fly Saliency Evaluator (OSE), an Accumulator, Digital/Analog WL Drivers (DWL/AWL), Digital Input Drivers (DIN), Analog Input Drivers (AIN) along with Digital-to-Analog Converters (DACs), a read/write IO (R/W IO), and a Controller. The  $64b \times 144b$  OSA-HCIM macro contains 8 Hybrid MAC Units (HMU), each of which comprises 144 Hybrid CIM Arrays (HCIMA), a Digital Adder Tree (DAT), a Normalization-and-Quantization Unit (N/Q), and a 3-bit SAR-ADC.

To perform *CIM* operations, OSA-HCIM executes both DCIM and ACIM within the HCIMA simultaneously. The MAC results of DCIM and ACIM, referred to as DMAC and AMAC, are then combined by shifting and adding them in the accumulator, resulting in the final multi-bit MAC output.

Further details regarding the functioning of DCIM and ACIM are provided below.

**Digital CIM (DCIM):** The digital input activations are dispatched to GBLBs in a bit-serial manner through DINs. Within each HCIMA, the digital circuitry carries out a bitwise multiplication of the stored weight and input, generating the Digital Output (DOUT). The DOUTs from the 144 columns are then aggregated by the Digital Adder Tree (DAT), yielding a 7-bit output DMAC.

**Analog CIM (ACIM):** Analog input activations of 1 to 4 bits are initially converted into analog voltages on GBLs utilizing DACs, and subsequently driven through AINs. The DAC is implemented via a switch matrix between reference voltages, enabling flexibility in bit-precision and supporting adaptable mapping of the analog MAC operations. In every HCIMA, the analog circuitry undertakes the multiplication of bit-serial weight and bit-parallel activation to produce AOUT. These AOUTs are summed using charge-sharing and converted to a 3-bit output AMAC via the SAR-ADC. Here, extremely low ADC precision is employed since ACIM is exclusively used for computing less significant data.

When carrying out the multi-bit MAC operations, the OSA-HCIM macro switches between two modes, namely, the Saliency Evaluation Mode and the Computing Mode.

**Saliency Evaluation Mode:** The purpose of this mode is to evaluate the saliency of the entire MAC operation utilizing few highest-order 1-bit MACs (Step 1 of Fig. 2(a)). In this mode, the DMACs are quantized to 3-bit via the Normalization-and-Quantization Unit (N/Q) and sent to the On-the-fly Saliency Evaluator (OSE). The OSE evaluates the saliency based on these quantized DMACs (Step 2 of Fig. 2(a)).

**Computing Mode:** Once the saliency is evaluated, the OSA-HCIM macro transitions to the Computing Mode. The role of this mode is to execute the remaining MAC operations based on the saliency score obtained from the Saliency Evaluation Mode (Step 3 of Fig. 2(a)). Since the OSE sets the digital-to-analog boundary to optimum  $B_{D/A}$ , the accuracy and resource

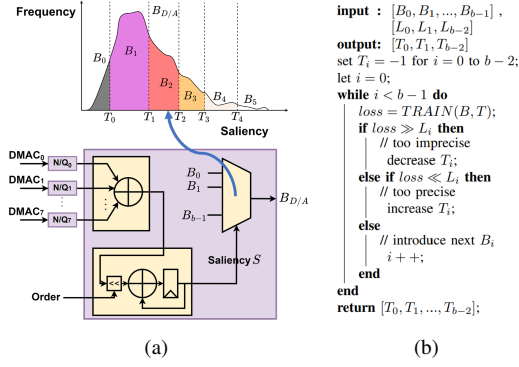


Fig. 3. The architecture of On-the-fly Saliency Evaluator (OSE). (b) The algorithm for determining  $B_{D/A}$  thresholds  $T$  for OSE.  $T$  can be pre-trained without degrading inference performance.

consumption will be balanced.

### B. Hybrid CIM Array (HCIMA)

Figure 2(c) illustrates the structure of our proposed Hybrid CIM Array (HCIMA), which is a novel CIM structure capable of *simultaneously* executing DCIM and ACIM, thereby effectively doubling the CIM throughput. The structure of HCIMA is composed of eight split-port 6T SRAMs, a pair of NMOS transistors (N0, N1) for conventional read/write operations during *RW* state, two PMOS transistors (P0, P1) responsible for the precharge of LBL and LBLB, and both digital ( $D_{MULT}$ ) and analog ( $A_{MULT}$ ) multipliers. Each HCIMA allows for the storage of either one 8-bit weight or two 4-bit weights. In *RW* state, the  $RW_{en}$  signal is elevated to activate N0 and N1. Both DWL and AWL from the target row are engaged for the standard SRAM read/write operations. Conversely, for *CIM* operation, the  $PCH$  signal is initially set down to enable P0/P1 to precharge LBLB/LBL.

Owing to the split-port readout scheme, different weights can be independently read on LBL and LBLB, thereby facilitating simultaneous digital and analog computations. Concurrently, GBLB and GBL transmit 1-bit inverted digital activation and 1~4-bit analog activation to  $D_{MULT}$  and  $A_{MULT}$  respectively, instigating the multiplication of weight and input activation. For example, in the case shown in Fig. 2(c), DWL7 and AWL0 are activated to readout  $Wb[7]$  on LBLB and  $W[0]$  on LBL. Meanwhile,  $Ab[7]$  and  $A[7:4]$  are sent via GBLB and GBL respectively. Subsequently, DOUT outputs the result of  $W[7] \times A[7]$ , and AOUT outputs the result of  $W[0] \times A[7:4]$ . The digital bit-serial multiplication of  $D_{MULT}$  is realized with a simple NOR gate, while the analog bit-parallel multiplication of  $A_{MULT}$  is realized with a transmission gate (T0), alongside an NMOS transistor (N2) for pull-down.

## V. SOFTWARE-HARDWARE CO-DESIGN: OSA-HCIM FRAMEWORK

The OSA-HCIM framework integrates the OSA precision configuration scheme of the software realm into the OSA-HCIM macro through the use of near-memory On-the-fly Saliency Evaluator (OSE). This framework also performs workload allocation to instigate DCIM and ACIM operations

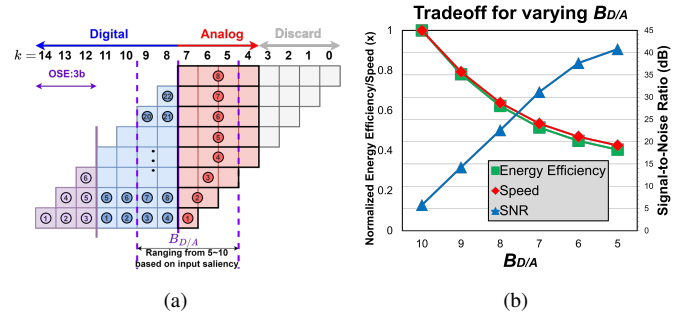


Fig. 4. (a) Workload allocation for DCIM and ACIM based on varying  $B_{D/A}$ . (b) Tradeoff between SNR, energy efficiency and execution speed under varying  $B_{D/A}$  for  $8b \times 8b$  MAC. The different Signal-to-Noise Ratio (SNR) and efficiency characteristics of the various  $B_{D/A}$  values meet the requirements for different types of input patterns, ensuring optimal performance across a range of scenarios.

effectively in HCIMA. The details of OSE and workload allocation method are discussed below.

### A. On-the-fly Saliency Evaluator (OSE)

The hardware implementation of the proposed On-the-fly Saliency Evaluator (OSE) is illustrated in Fig.3(a). OSE plays a critical role in our framework — it estimates the saliency and determines the suitable  $B_{D/A}$  of the multi-bit MAC operation from the high-order 1-bit MAC results. To achieve this, OSE first normalizes and quantizes the high-order 1-bit MAC result, denoted  $DMAC_i$  in Fig.3(a), of each channel. Then, these values are summed up and accumulated across cycles to obtain the saliency value  $S$ . Based on the calculated saliency  $S$ , the OSE determines a suitable  $B_{D/A}$  from a candidate list  $B = [B_0, B_1, \dots, B_{b-1}]$ . This is done by comparing  $S$  with a set of thresholds  $T = [T_0, T_1, T_{b-2}]$ . This process is visualized in the histogram in Fig. 3(a).

The thresholds  $T$  are determined based on the algorithm illustrated in Fig. 3(b). The algorithm is supplied with  $B$  and a set of training loss constraints  $L = [L_0, L_1, \dots, L_{b-2}]$ . Through an iterative process that explores the threshold  $T_i$  within the boundaries  $B_i$  and  $B_{i+1}$  to match the loss constraint  $L_i$ , a desired set of thresholds  $T$  is obtained. These thresholds are pre-trained, hence, they do not incur any additional overhead during the inference. Importantly, the loss constraints  $L$  are specified by the user, allowing for customization of the desired tradeoff between accuracy and efficiency. This adaptability renders the framework compatible with a wide range of tasks.

### B. Workload Allocation

In this section, we propose a workload allocation to maximize the utilization of our hybrid CIM with digital and analog concurrent computation. Figure 4(a) illustrates the workload allocation for an  $8b \times 8b$  MAC. Based on the OSA precision configuration scheme, each 1-bit MAC is partitioned into three sections: *digital*, *analog*, or *discard*. *Digital mode* MACs are scheduled once per cycle using DCIM. Conversely, *analog* 1-bit MACs with the same weight are calculated simultaneously in a bit-parallel manner using ACIM. Here, the variable bit-width is accommodated by the variable-precision DAC. Then,



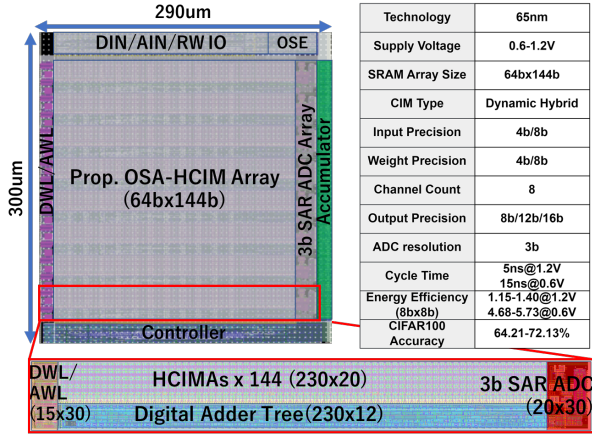


Fig. 5. The layout of OSA-HCIM and its summary.

input activations are routed through GBLB/GBL, while weight access is facilitated by configuring DWL/AWL.

Our scheme demonstrates its capability in providing a diverse selection of operating points with distinct precision and efficiency characteristics. As illustrated in Fig. 4(b), each value of  $B_{D/A}$  ranging from 5 to 10 represent a valuable operating point with distinct characteristics in terms of SNR and efficiency. Compared to prior works that only utilize dual-precision for Saliency-Aware computing [13], [14], [17], our methodology provides a significantly broader range of options to tailor to the varying saliency of each input.

Based on the allocation result, DCIM and ACIM compute the different assigned workloads concurrently, attaining an improved throughput. One risk associated with the allocation scheme is the possibility of unbalanced workload between DCIM and ACIM, originated from the inherent throughput discrepancy and the variable  $B_{D/A}$  value. To counteract this, DCIMs can be operated at a higher clock frequency, leveraging the fact that DAT has twice lower latency compared to ADC counterparts, since the SAR ADC requires 3 cycles to complete the conversion.

## VI. SIMULATION RESULTS

We implemented the  $64b \times 144b$  OSA-HCIM on a 65nm CMOS process. Figure 5 shows the layout and the summary of OSA-HCIM, and the power and area breakdowns of OSA-HCIM is shown in Fig. 6. A significant advantage of our design is the OSE component—despite its critical role in facilitating the OSA precision configuration scheme, it incurs only a minimal overhead of 1% in power and 1% in area, largely due to the compressed  $DMAC$  bandwidth and cost amortization across 8 HMUs. Furthermore, unlike previous ACIM designs where the ADC was a primary power and area bottleneck, OSA-HCIM allows the utilization of low-precision 3-bit SAR ADC, which accounts for merely 17% of the total power and 6% of the total area.

The  $B_{D/A}$  maps of various hidden layers processing a horse image using ResNet20 are shown in Fig. 7(a). As observed in these layers, the majority of pixels pertaining

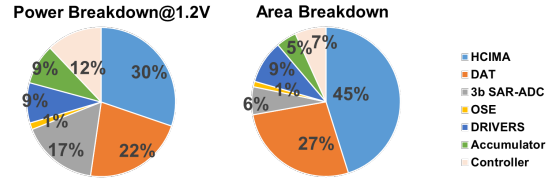


Fig. 6. The power and area breakdowns of OSA-HCIM. OSE incurs modest overhead thanks to the compressed input bandwidth and cost amortization across HMUs.

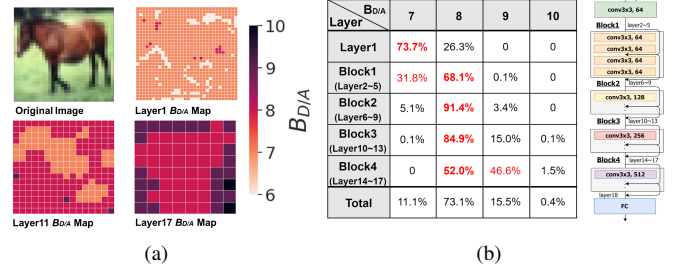


Fig. 7. (a)  $B_{D/A}$  maps of hidden layers for a horse image using ResNet20. OSE automatically maps high-precision computing settings to pixels pertaining the horse, and low-precision settings are assigned to background pixels. (b) The proportion of each  $B_{D/A}$  across all CONV blocks of ResNet18 tested on CIFAR100 dataset. OSE effectively adapts to the optimum precision requirements across different layers.

to the horse are detected and assigned with relatively high-precision  $B_{D/A}$  settings, while background pixels are assigned to low-precision settings. These observations highlight the effectiveness of OSE in identifying salient pixels and adapting the fine-grained precision accordingly. The adaptability of OSA-HCIM to saliency across layers is further evidenced in the analysis of  $B_{D/A}$  usage in ResNet18 for the CIFAR100 dataset, depicted in Fig. 7(b). With the progression into deeper layers, there is a noticeable increase in the utilization of low-precision settings, effectively saving computational resources. The exceptional level of adaptability exhibited by OSA-HCIM, which exceeds that of previous works [6], achieves an optimal balance between accuracy and overall efficiency.

Figure 8 presents the correlation between CIFAR100 accuracy under ResNet18 and energy efficiency of DCIM, HCIM (without OSA), and OSA-HCIM. HCIM, which combines CIM of both analog and digital domains, achieves a 1.56x improvement in energy efficiency with  $< 2\%$  accuracy loss. OSA-HCIM, taking further advantage of input saliency, achieves an additional 1.25x efficiency boost, cumulating to a 1.95x total improvement compared to DCIM. In addition, OSA-HCIM offers a diverse set of operating points with unique accuracy-efficiency tradeoffs by adjusting the constraints  $L$  in the threshold-finding algorithm in OSE, as discussed in Section V. It is capable of either obtaining 0.1% accuracy drop or achieving high energy efficiency of 5.79 TOPS/W. The flexibility afforded by the configurability of OSE empowers OSA-HCIM to address a variety of real-world tasks, each with its own precision requirements. For instance, OSA-HCIM can achieve high accuracy even applied to more challenging datasets, such as ImageNet, since the algorithm

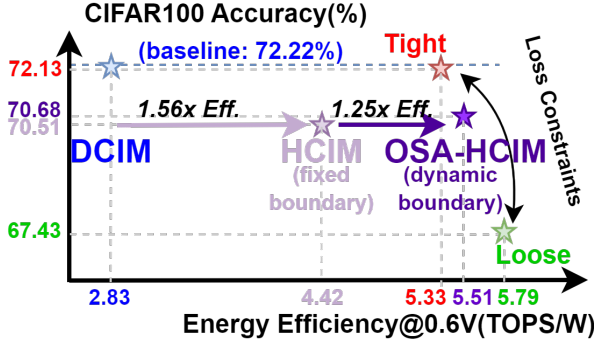


Fig. 8. By adjusting the loss constraints  $L$ , OSA-HCIM can adapt its operation to prioritize either higher accuracy or higher energy efficiency, depending on the specific requirements.

in Fig. 3(b) autonomously explores thresholds that prioritize higher precision boundaries, with some tradeoff in efficiency.

Table I presents the comparison with state-of-the-art designs. Notably, OSA-HCIM introduces the concept of Saliency-Aware Computing to CIM, enabling adaptive processing reflecting the input patterns. It is also the first CIM implementation to feature dynamic configuration of the digital-to-analog boundary, offering significant flexibility in terms of computing precision. Unlike conventional ACIMs, which often struggle to achieve high accuracy with large-scale datasets (e.g. CIFAR100, ImageNet), OSA-HCIM attains accuracy close to those of DCIM counterparts. Moreover, OSA-HCIM delivers an energy efficiency of 5.33-5.79 TOPS/W, which is competitive against SRAM CIM designs when normalized to the 65nm process under CIFAR100 dataset.

## VII. CONCLUSIONS

Although CIM has been refined for general-purpose MAC operations, its efficiency suffers due to the lack of dynamic precision computing, which limits its adaptability to the saliency of different inputs. In response to this challenge, we introduced OSA-HCIM, a hybrid CIM architecture developed through software-hardware co-design. The OSA precision configuration scheme enables the capability for online evaluation of input saliency, thus providing an array of precision settings tailored to different input patterns, enhancing computational accuracy and efficiency. Furthermore, HCIMA, a CIM topology that enables simultaneous computation in the digital and analog domains, increases computational flexibility and boosts throughput. The OSA-HCIM framework, bridging the OSA precision configuration scheme with HCIMA, exhibits remarkable versatility, effectively tackling tasks that demand varying precision requirements. The proposed design, which is implemented using 65nm CMOS process, achieves a high energy efficiency of 5.33-5.79 TOPS/W while maintaining robust inference accuracy tested on the CIFAR100 dataset.

## ACKNOWLEDGEMENTS

This research was supported in part by the JST CREST JPMJCR21D2, JSPS Kakenhi 23H00467, Futaba Foundation, Asahi Glass Foundation, and

the Telecommunications Advancement Foundation.

TABLE I  
COMPARISON WITH STATE-OF-THE-ART SRAM CIM MACROS

	ICCAD '22 [7]	ISSCC '21 [4]	MCSoc '22 [8]	This Work
<b>Tech. (nm)</b>	28	22	22	<b>65</b>
<b>CIM Type</b>	Analog	Digital	Fixed Hybrid	<b>Dynamic Hybrid</b>
<b>Input Prec.</b>	4b	1-8b	1b	<b>4/8b</b>
<b>Weight Prec.</b>	8b	4/8/12/16b	8b	<b>4/8b</b>
<b>Supply (V)</b>	0.6-1.2	0.72	0.8	<b>0.6-1.2</b>
<b>Array Size</b>	256x64	256x256	64x96	<b>64x144</b>
<b>CIFAR100 Acc. (drop)<sup>c</sup></b>	65.8% (0.5%)	- (0%)	71.92% (4.17%)	<b>67.4~72.1% (4.8~0.1%)</b>
<b>ImageNet Acc. (drop)</b>	- (-)	- (0%)	- (-)	<b>65.2~70.8% (6.3~0.8%)</b>
<b>Energy Eff. (TOPS/W)<sup>a</sup></b>	5.7-22.9	24.7	6.98-11.0	<b>5.33~5.79<sup>d</sup> @ CIFAR100 3.83~4.66<sup>d</sup> @ ImageNet</b>
<b>Norm. Energy Eff. (TOPS/W)<sup>b</sup></b>	1.06-4.25	2.83	0.80-1.26	
<b>Saliency-Aware</b>	No	No	No	<b>Yes</b>

<sup>a</sup>Normalized to 8b × 8b MACs (1 MAC is 20Ps).

<sup>b</sup>Normalized to 65nm process.

<sup>c</sup>Compared to the baseline accuracy reported in each work.

<sup>d</sup>Simulated under 0.6V supply voltage

## REFERENCES

- [1] J. Yue *et al.*, “15.2 a 2.75-to-75.9 tops/w computing-in-memory nn processor supporting set-associate block-wise zero skipping and ping-pong cim with simultaneous computation and weight updating,” in *ISSCC*, vol. 64. IEEE, 2021, pp. 238–240.
- [2] —, “A 28nm 16.9-300tops/w computing-in-memory processor supporting floating-point nn inference/training with intensive-cim sparse-digital architecture,” in *ISSCC*. IEEE, 2023, pp. 1–3.
- [3] B. Wang *et al.*, “A 28nm horizontal-weight-shift and vertical-feature-shift-based separate-w1 6t-sram computation-in-memory unit-macro for edge depthwise neural-networks,” in *ISSCC*. IEEE, 2023, pp. 134–136.
- [4] Y.-D. Chih *et al.*, “16.4 an 89tops/w and 16.3 tops/mm<sup>2</sup> all-digital sram-based full-precision compute-in memory macro in 22nm for machine-learning edge applications,” in *ISSCC*, vol. 64. IEEE, 2021, pp. 252–254.
- [5] C.-F. Lee *et al.*, “A 12nm 121-tops/w 41.6-tops/mm<sup>2</sup> all digital full precision sram-based compute-in-memory with configurable bit-width for ai edge applications,” in *VLSI*. IEEE, 2022, pp. 24–25.
- [6] K. Lee *et al.*, “A charge-sharing based 8t sram in-memory computing for edge dnn acceleration,” in *DAC*. IEEE, 2021, pp. 739–744.
- [7] —, “Low-cost 7t-sram compute-in-memory design based on bit-line charge-sharing based analog-to-digital conversion,” in *ICCAD*, 2022, pp. 1–8.
- [8] J. Chen *et al.*, “A charge-digital hybrid compute-in-memory macro with full precision 8-bit multiply-accumulation for edge computing devices,” in *MCSoc*. IEEE, 2022, pp. 153–158.
- [9] P.-C. Wu *et al.*, “A 22nm 832kb hybrid-domain floating-point sram in-memory-compute macro with 16.2-70.2 tflops/w for high-accuracy ai-edge devices,” in *ISSCC*. IEEE, 2023, pp. 126–128.
- [10] H. Hadizadeh *et al.*, “Saliency-aware video compression,” *TIP*, vol. 23, no. 1, pp. 19–33, 2013.
- [11] N. G. Sadaka *et al.*, “Efficient super-resolution driven by saliency selectivity,” in *ICIP*. IEEE, 2011, pp. 1197–1200.
- [12] Z. Ren *et al.*, “Region-based saliency detection and its application in object recognition,” *TCSVT*, vol. 24, no. 5, pp. 769–779, 2013.
- [13] Y. Zhang *et al.*, “Precision gating: Improving neural network efficiency with dynamic dual-precision activations,” *arXiv preprint arXiv:2002.07136*, 2020.
- [14] Z. Song *et al.*, “Drq: Dynamic region-based quantization for deep neural network acceleration,” in *ISCA*, 2020, pp. 1010–1021.
- [15] M. R. H. Rashed *et al.*, “Hybrid analog-digital in-memory computing,” in *ICCAD*. IEEE, 2021, pp. 1–9.
- [16] C.-Y. Yao *et al.*, “A fully bit-flexible computation in memory macro using multi-functional computing bit cell and embedded input sparsity sensing,” *JSSC*, vol. 58, no. 5, pp. 1487–1495, 2023.
- [17] L. Liu *et al.*, “Duet: Boosting deep neural network efficiency on dual-module architecture,” in *MICRO*. IEEE, 2020, pp. 738–750.