RESEARCH HIGHLIGHTS

Journal of Semiconductors
(2023) 44, 050203
doi: 10.1088/1674-4926/44/5/050203

Beyond convolutional neural networks computing: New trends on ISSCC 2023 machine learning chips

Chen Mu[1], Jiapei Zheng[1], and Chixiao Chen[1, 2, †]

[1]State Key Laboratory of Integrated Chips and Systems, Frontier Institute of Chips and Systems, Fudan University, Shanghai 200433, China
[2]Qizhi Institute, Shanghai 200232, China

Citation: C Mu, J P Zheng, and C X Chen, Beyond convolutional neural networks computing: New trends on ISSCC 2023 machine learning chips[J]. *J. Semicond.*, 2023, 44(5), 050203. https://doi.org/10.1088/1674-4926/44/5/050203

Machine learning (ML) domain specific architectures (DSA) and chips have been prevailed in the past few years. These custom DSA designs outperform conventional general purposed architectures in terms of energy efficiency, processing latency and performance scalability. The data intensive nature of ML requires large amounts of processing power and memory access. Data flow architectures, reconfigurable near-and-in-memory circuits were proposed for convolutional neural networks (CNNs), resulting in tremendous power reduction. Futhermore, complex intelligent tasks on edge, such as autonomous vehicles and robotics, are likely to require real-time processing of data. The custom NN computing pipeline improves overall throughput and latency. In addition, traditional processors may not be able to keep up with the demand from deeper and deeper networks. Domain-specific multi-core architectures were designed to scale with the size and complexity of the networks, facilitating emerging algorithms.

However, ML algorithm is a rapidly evolving research area, where new techniques and approaches are constantly explored to improve performance on a variety of tasks. One important trend in recent years has been the exploration of deep learning architectures beyond convolutional neural networks (CNNs), such as self-attention based transformers, reinforcement learning, spike based neuromophic computing. As a result, there is a growing demand for specialized ML chips that can accelerate a wide range of models beyond CNNs. Under the trend, there are couple of specialized chips designed for the merging AI tasks and architectures on ISSCC 2023. In this review, the authors categorize these chips into four trends, and briefly go through the representative works, illustrated in Fig. 1.

Trend I: Compute-in-memory designs supporting floating point operation

In general, CNN chips allow low bit-width quantized integer computing, such as binary, ternary, or INT4 or INT8. The resolution is affordable for previous Compute-in-memory (CIM) designs. However, when confronted with more complex tasks or network models, these low-precision chips tend to perform worse than floating-point ones. Recognizing that current CIM chips do not effectively support floating-point numbers, researchers presented CIM circuit and systems supporting floating-point operations. National Tsinghua University[1], Southeast University[2], and the Institute of Microelectronics of the Chinese Academy of Sciences[3] presented their proposed floating-point CIM chips at ISSCC 23. The increasingly effective support for floating-point operations shows the potential of CIM chips in various scenarios. Wu from National Tsinghua University[1] proposed a hybrid-domain floating-point SRAM In-Memory-Compute Macro that leverages the benefits of time, digital, and analog domains. A time-domain-based exponent summation array, maximum product exponent finder, and product exponent difference generator were proposed for energy-efficient exponent calculation. The FP SRAM-IMC supports BF16 input and weight, FP32 output, and achieves 70.2 TFLOPS/W energy efficiency. Guo from Southeast University[2] proposed a digital-domain FP-MAC SRAM-CIM structure, specifically comprising (1) a cell array with double-bit cells, each of which stored two continuous weight bits, (2) floating-point computing units with high-bit full-precision multiply cell and low-bit approximate-calculation multiply cell to reduce internal bandwidth and area, and (3) a CIM-macro architecture with FP processing circuits to support both FP-MAC and INT-MAC operations. This FP-MAC supports input, weight, and output of BF16, and achieves an energy efficiency of 31.6 TFLOPS/W and an area efficiency of 2.05 TFLOPS/mm^2.

Yue from Institute of Microelectronics of the Chinese Academy of Sciences[3] proposed a processing approach for floating-point data that divides it into dense CIM and sparse digital parts. For the dense part, an FP-to-INT CIM workflow was proposed to reduce execution cycles and improve efficiency. For the sparse part, a flexible sparse digital core was proposed to encode sparse activation/weight data. This helps CIM cores achieve high energy efficiency and precision. The chip achieves 17.2–91.3 TOPS/W@FP16 macro energy efficiency for inference tasks.

Trend II: Specialized ML chips for transformer

Recently, the large natural language learning model and visual model based on transformer have become a hot research topic. The emergence of ChatGPT has also attracted unprecedented attention. However, the large storage and computing requirements involved in the transformer network also pose challenges to its specific hardware accelerator design. In view of this, scholars from different universities on ISSCC 2023 reported various optimized hardware designs based on computing in memory or near memory computing methods for different challenges. Tu from Tsinghua University proposed a digital CIM-based MulT model accelerator MulTCIM for Multimodal transformer, which solves the prob-

Correspondence to: C X Chen, cxchen@fudan.edu.cn
Received 25 MARCH 2023.
©2023 Chinese Institute of Electronics

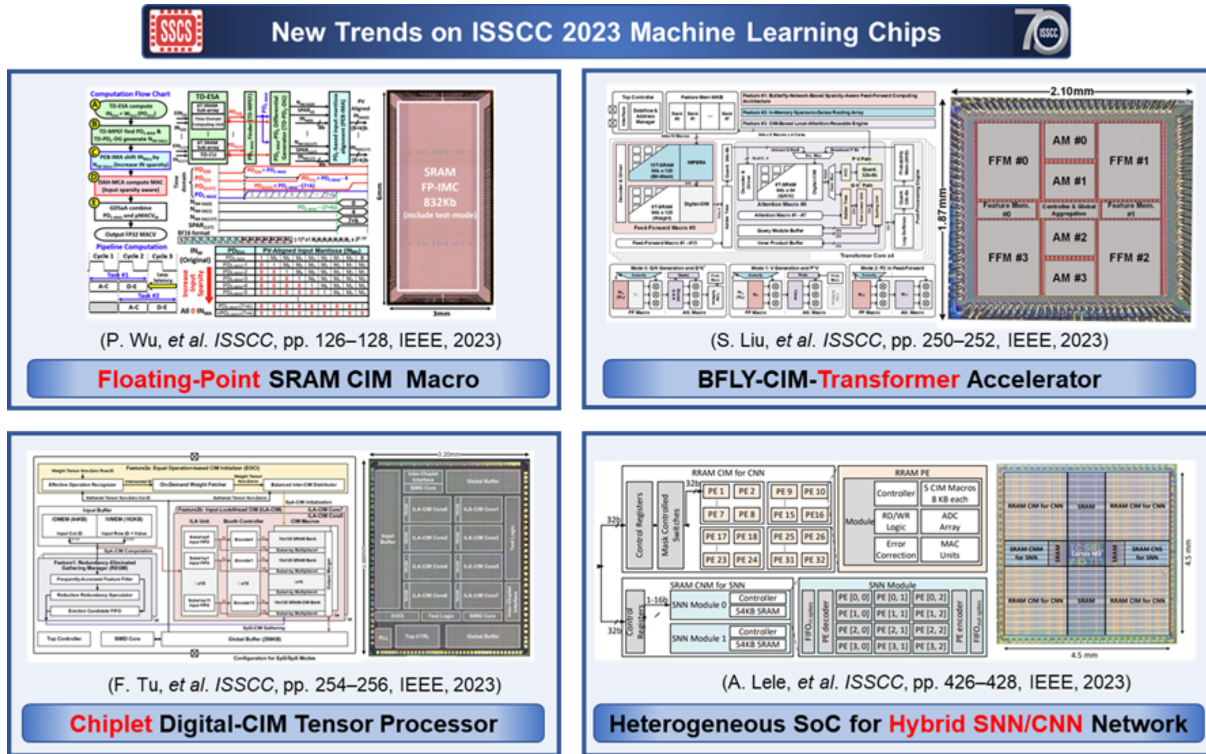

Fig. 1. (Color online) New trends on ISSCC 2023 machine learning chips.

lems of low CIM utilization in long reuse distance of attention sparsity, high latency in cross-modal switch, and more variance in effective bitwidth for the same group of inputs in a cim macro[4]. This work obtain 9.47× speedup and 8.11× energy savings on ViLBERT-base's attention layers, achieving 6.54× speedup and 5.61× energy savings on the entire model. Liu from Fudan University proposed an efficient transformer accelerator, featuring custom in-memory routing and computing macros for the unstructured sparsity in transformer networks to solve the challenge for CIM-based accelerators that how to handle unstructured pruned NNs, while maintaining high efficiency[5]. Evaluation is conducted on a 12-layer 8-head adaptive-attention-span transformer model with 1024 input tokens for the Enwik-8 and Text-8 tasks. This work achieved 53.83 TOPS/W energy efficiency and 0.85 TOPS/mm^2 area efficiency in 28 nm CMOS technology. Tambe from Harvard University also presented a 4.60 mm^2 sparse transformer processor (STP) that efficiently accelerates transformer workloads by tailoring its latency and energy expenditures according to the complexity of the input query it processing[6]. A fine-grained sentence-level power management scheme is interesting that opportunistically scales the accelerator's supply voltage and clock frequency while meeting an application's end-to-end latency target. This work produced an energy efficiency range of 3.0–8.24 TFLOPS/W (FP8) and 6.61–18.1 TFLOPS/W (FP4) in 12 nm CMOS technology.

Trend III: Heterogeneous SOCs for hybrid SNN/CNN networks

Spike neural networks (SNNs) have unique advantages in perceived speed with event-driven characteristics, throughput and low power consumption, but SNNs cannot achieve the same accuracy as convolution neural networks (CNNs) through backpropagation training. Compared with SNNs, CNNs require more computing resources, energy consumption, and computing delay to achieve higher recognition accuracy. In this year's ISSCC, heterogeneous architectures combining SNN and CNN are proposed to realize AI intelligent hardware with high throughput, low power consumption, and high recognition accuracy. Chang from Georgia Institute of Technology presents a fully-programmable heterogeneous ARM Cortex-based SoC with an in-memory low-power RRAM-based CNN and a near-memory high-speed SRAM-based SNN in a hybrid architecture with applications in high-speed target identification and tracking[7]. In order to reduce the energy consumption of CNN inferencing process, this work features a two-level power gating for RRAM-based engine indicating a 78% power reduction.This fused system achieves small accuracy degradation compared to CNN-only with 10× higher throughput and lower average power consumption for two CNN inferences/second. Kim from Korea Advanced Institute of Science and Technology proposed a Complementary Deep Neural Network (C-DNN) processor combining CNNs and SNNs to take advantage of their respective strong points[8]. The heterogeneous processor can be allocated between SNN and CNN core according to different tasks to realize low-power inferencing, and can also assist CNN to complete more efficient back propagation through SNN during training process. This work shows 77.1% accuracy for ImageNet with state-of-the-art energy efficiency of 24.5 TOPS/W (in ResNet-50). In addition, in order to optimize the large energy and time consumption caused by training on SNN tablets, Zhang from Tsinghua University presented ANP-I, an asynchronous SNN processor enabling on-chip training and consuming only 0.5-15% of the inference energy per sample[9]. This work achieves 2150× and 56× on-chip training energy savings on hand gesture recognition and KWS tasks with 4.7% and 1.9% accuracy improvement respectively over SOTA.

Trend IV: Multiple-chiplet integrated chips for high performance computing

Dr. Lisa T. Su, the chair and chief executive officer of AMD, deliver a plenary speech on the topic of *Innovation For the Next Decade of Compute Efficiency*[10]. Advanced packaging technology was emphasized as being increasingly attractive for use in modular chiplet architectures, where only the most advanced nodes are used for the most compute-heavy IP that gains the most from using the latest technology. Recently, state-of-the-arts processors are likly to use 2.5D/3D integration technology. At the conference, AMD[11], Samsung[12], Tesla[13], and Tsinghua University[14] shared their progress in the integration of the chiplet system. Munger from AMD[11] proposed mainstream client products based on "Zen 4" architecture that combines the client IOD with 1 or 2 "Zen 4" CCD chiplets for 16-, 12-, 8-, and 6-core products. Its L3 cache supports the second generation AMD 3D V-cache, which extends it from 32 MB to 96 MB per CCX. The overhead area to support V-cache was reduced by 40% in the second-generation implementation. Seong from Samsung[12] implemented a die-to-die chiplet based on 2.5D packaging technology for die-to-die communication compatible with UCIe specification. The chiplet features an NRZ Single-Ended Transceiver with equalization schemes and training techniques operating at 32 Gb/s/wire in 4 nm FinFET CMOS technology. It shows 0.44 pJ/b energy efficiency and 8 Tb/s/mm bezel-front bandwidth. Fischer from Tesla[13] introduced their D1 chip as the ML training processor in the DOJO exa-scale computer system. The D1 processor achieves 362 TFlops of BFP16/CFP8 performance and is implemented in TSMC 7 nm process with advanced fan-out packaging. It has 144 custom 112 Gbps SerDes lanes on each chip edge for die-die communication, providing a bidirectional BW of 4 TB/s. Tu from Tsinghua University[14] proposed TensorCIM, a scalable system solution for beyond-NN that uses a multi-chip-module to provide increased computing power and memory capacity while reducing manufacturing costs. TensorCIM combines digital CIM with reconfigurable technology to dynamically switch between sparse tensor aggregation and sparse neural network computing and maintain high resource utilization. It achieves a sparse tensor aggregation efficiency of 3.7 nJ/gather and a sparse FP32 tensor algebraic efficiency of 8.3 TFLOPS/W.

In conclusion, this year's ISSCC suggests the trends of ML chips, towards increased efficiency and flexibility, with a focus on accommodating a wider range of machine learning algorithms beyond just convolutional neural networks (CNNs).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2022YFB4500100).

References

[1] Wu P C, Su J W, Hong L Y, et al. A 22nm 832Kb hybrid-domain floating-point SRAM in-memory-compute macro with 16.2-70.2TFLOPS/W for high-accuracy AI-edge devices. 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, 126

[2] Guo A, Si X, Chen X, et al. A 28nm 64-kb 31.6-TFLOPS/W digital-domain floating-point-computing-unit and double-bit 6T-SRAM computing-in-memory macro for floating-point CNNs. 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, 128

[3] Yue J S, He C J, Wang Z, et al. A 28nm 16.9-300TOPS/W computing-in-memory processor supporting floating-point NN inference/training with intensive-CIM sparse-digital architecture. 2023 IEEE International Solid- State Circuits Conference (ISSCC), 2023, 252

[4] Tu F B, Wu Z H, Wang Y Q, et al. MulTCIM: A 28nm 2.24 μj/token attention-token-bit hybrid sparse digital CIM-based accelerator for multimodal transformers. 2023 IEEE International Solid- State Circuits Conference (ISSCC), 2023, 248

[5] Liu S, Li P, Zhang J, et al. A 28nm 53.8TOPS/W 8b sparse transformer accelerator with in-memory butterfly zero skipper for unstructured-pruned NN and CIM-based local-attention-reusable engine. 2023 IEEE International Solid- State Circuits Conference (ISSCC), 2023, 250

[6] Tambe T, Zhang J, Hooper C, et al. A 12nm 18.1TFLOPs/W sparse transformer processor with entropy-based early exit, mixed-precision predication and fine-grained power management. 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, 342

[7] Chang M Y, Lele A S, Spetalnick S D, et al. A 73.53TOPS/W 14.74TOPS heterogeneous RRAM In-memory and SRAM near-memory SoC for hybrid frame and event-based target tracking. 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, 426

[8] Kim S, Kim S, Hong S, et al. C-DNN: A 24.5-85.8TOPS/W complementary-deep-neural-network processor with heterogeneous CNN/SNN core architecture and forward-gradient-based sparsity generation. 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, 334

[9] Zhang J L, Huo D X, Zhang J, et al. ANP-I: A 28nm 1.5pJ/SOP asynchronous spiking neural network processor enabling sub-O.1 μJ/sample on-chip learning for edge-AI applications. 2023 IEEE International Solid- State Circuits Conference (ISSCC), 2023, 21

[10] Su L S, Naffziger S. Innovation for the next decade of compute efficiency. 2023 IEEE International Solid- State Circuits Conference (ISSCC), 2023, 8

[11] Munger B, Wilcox K, Sniderman J, et al. “Zen 4”: The AMD 5nm 5.7GHz x86-64 microprocessor core. 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, 38

[12] Seong K, Park D, Bae G, et al. A 4nm 32Gb/s 8Tb/s/mm Die-to-Die chiplet using NRZ single-ended transceiver with equalization schemes and training techniques. 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, 114

[13] Fischer T C, Nivarti A K, Ramachandran R, et al. 9.1 D1: A 7nm ML training processor with wave clock distribution. 2023 IEEE International Solid- State Circuits Conference (ISSCC), 2023, 8

[14] Tu F B, Wang Y Q, Wu Z H, et al. 16.4 TensorCIM: A 28nm 3.7nJ/gather and 8.3TFLOPS/W FP32 digital-CIM tensor processor for MCM-CIM-based beyond-NN acceleration. 2023 IEEE International Solid- State Circuits Conference (ISSCC), 2023, 254

Chen Mu received the B.S. degree from Southeast University, Nanjing, China, in 2020. He is currently pursuing Ph.D. degree with State Key Laboratory of Integrated Chips and Systems, Frontier Institute of Chips and Systems, Fudan University, Shanghai, China. His current research interests include computing-in-memory, high-efficient AI accelerator and chiplets.

Jiapei Zheng received the B.S. degree in microelectronics from Fudan University in 2021. He is currently pursuing the Ph.D. degree with State Key Laboratory of Integrated Chips and Systems, Frontier Institute of Chips and Systems, Fudan University, Shanghai, China. His research interests include custom intelligent software-hardware co-designs and intelligent 3-D vision accelerators.

Chixiao Chen received the B.S. and Ph.D. degrees in Microelectronics from Fudan University, Shanghai, China in 2010 and 2015, respectively. From 2016 to 2018, he was a post-doctoral research associate with the University of Washington, Seattle. Since 2019, he has been with Fudan University, Shanghai, China as an Assistant Professor, where he is currently an Associate Professor. He is also an adjunct research associate with the National Key Laboratory of Integrated Chips and Systems, Fudan University. His research interest includes mixed-signal integrated circuit design and custom intelligent software-hardware co-designs.