Journal of Semiconductors
JOS iopscience.iop.org/jos
www.jos.ac.cn

Towards efficient generative AI and beyond-AI computing: New trends on ISSCC 2024 machine learning accelerators

Bohan Yang, Jia Chen, and Fengbin Tu

Citation: B H Yang, J Chen, and F B Tu, Towards efficient generative AI and beyond–AI computing: New trends on ISSCC 2024 machine learning accelerators[J]. *J. Semicond.*, 2024, 45(4).

View online: https://doi.org/10.1088/1674–4926/45/4/040204

RESEARCH HIGHLIGHTS

Journal of Semiconductors
(2024) 45, 040204
doi: 10.1088/1674-4926/45/4/040204

Towards efficient generative AI and beyond-AI computing: New trends on ISSCC 2024 machine learning accelerators

Bohan Yang[1, 3], Jia Chen[1, 2], and Fengbin Tu[1, 2, †]

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[2]AI Chip Center for Emerging Smart Systems, The Hong Kong University of Science and Technology, Hong Kong, China
[3]School of the Gifted Young, University of Science and Technology of China, Hefei 230026, China

Citation: B H Yang, J Chen, and F B Tu, Towards efficient generative AI and beyond-AI computing: New trends on ISSCC 2024 machine learning accelerators[J]. *J. Semicond.*, 2024, 45(4), 040204. https://doi.org/10.1088/1674-4926/45/4/040204

Compared to the last decade when the convolution neural network (CNN) dominated the research field, machine learning (ML) algorithms have reached a pivotal moment called the generative artificial intelligence (AI) era. With the emergence of large-scale foundation models[1], such as large multimodal model (LMM) GPT-4[2] and text-to-image generative model DALL·E[3], advanced ML accelerators should address challenging scaling problems from both compute and memory. Meanwhile, with the growing demand for general intelligence on diverse application scenarios (robotics, automobiles, digital economy, manufacturing, etc.), designing integrated smart systems with intelligent perception, processing, planning, and action capabilities is a trend for future AI platforms. Besides the AI processing engine, an integrated smart system also requires domain-specific architectures (DSA) for beyond-AI computing.

Last year, Prof. Chixiao Chen's group from Fudan University summarized trends for ML chips beyond CNN computing in the "Chip Olympiad" ISSCC 2023[4]. In this survey, we take a deep look into ISSCC 2024 and observe four research trends toward efficient generative AI (**ML chips for generative AI, computing-in-memory (CIM) innovation from circuits to systems**) and beyond-AI computing (**DSA for embedded vision processors, DSA for solver accelerators**), as illustrated in Fig. 1. We believe these remarkable trends will lead to more AI software and hardware innovations from academia and industry in the near future.

Trend 1: ML chips for generative AI

Since 2022, generative AI models have demonstrated remarkable capabilities in creating new data samples based on the probabilistic model learned from huge datasets. By generating high-quality and coherent text, images, or even control signals, generative AI has the potential to revolutionize any field with its creative sample generation skills. In order to power this new era of AI platforms with high computing and memory efficiency, there is a trend of ML chips for generative AI observed in different events of ISSCC 2024, including the plenary speech, main conference, and forum.

In ISSCC 2024, Dr. Jonah Alben, senior vice president of NVIDIA, delivers a plenary speech about *Computing in the Era of Generative AI*[5]. He shows the possibility of alleviating the gap between AI algorithms and underlying computing systems by letting AI help us design chips because they can find more optimized architecture and circuits. In the main conference, AMD proposes MI300 series processors for generative AI and high-performance computing (HPC) workloads with modular chiplet package and cutting-edge HBM3 memory which has up to 192 GB capacity and 5.3 TB/s peak theoretical bandwidth[6]. Guo *et al*. from Tsinghua University propose the first heterogenous CIM-based accelerator for image-generative diffusion models[7], which leverages pixel similarity between denoising iterations to apply mixed quantization. They propose novel methods in bit-parallel CIM by booth-8 multiplier, balanced integer/floating-point (INT/FP) processing latency by exponent processing acceleration, and support FP sparsity by in-memory redundancy search. Kim *et al*. from KAIST target hybrid SNN-transformer inference on edge[8]. Spiking neural network (SNN) has supremacy in low power and high efficiency with low-bitwidth accumulation-only discrete operations and can be integrated with transformers. They realize an ultra-low-power on-device inference system by hybrid multiplication/accumulation units, speculative decoding, and implicit weight generation, reducing external memory access (EMA) by 74%–81%. In addition, a special forum named *Energy-Efficient AI-Computing Systems for Large-Language Models* shares more practical thoughts about large language model (LLM) computing systems[9]. Georgia Institute of Technology, NVIDIA, Intel, Google, KAIST, Samsung, Axelera AI, and MediaTek introduce their latest research over LLM training and inference in both cloud and edge.

Trend 2: CIM innovation from circuits to systems

Computing-in-memory (CIM) technology, as a promising way to break the memory wall of traditional Von Neumann architecture, has gained significant popularity in the integrated circuit and computer architecture communities. As complex ML algorithms advance, new challenges for CIM design are widely investigated. In ISSCC 2024, we have noticed many CIM innovations from energy-efficient circuits, floating-point (FP) support, to system integration.

The lookup table (LUT)-based digital CIM (DCIM) scheme presents a promising enhancement in the compute density of traditional single-bit-input DCIM design. TSMC reports an advanced 3 nm-node DCIM macro with a parallel multiply-accumulation (MAC) architecture based on LUTs[10]. He *et al*. from Tsinghua University also propose an eDRAM-LUT-based DCIM macro for compute and memory-intensive applications[11].

Correspondence to: F B Tu, fengbintu@ust.hk
Received 8 MARCH 2024; Revised 26 MARCH 2024.

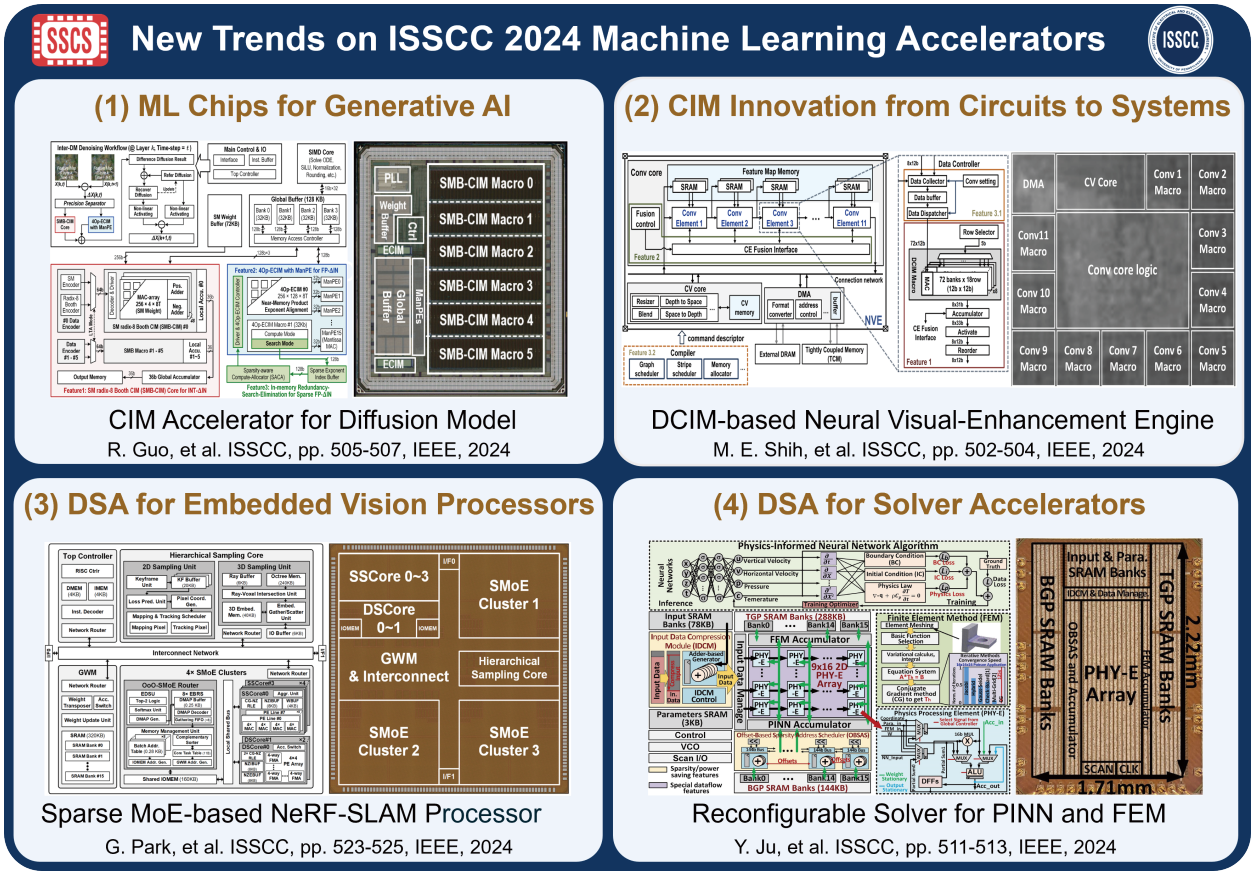

Fig. 1. (Color online) New trends on ISSCC 2024 machine learning accelerators.

Hybrid CIM is another appealing direction in ISSCC 2024 to integrate the advantages of analog and digital schemes. Guo *et al*. from Southeast University design a lightning-like analog/digital hybrid CIM structure to capitalize on both the high energy efficiency of analog CIM (ACIM) and the robust accuracy of DCIM[12]. Wang *et al*. from the Institute of Microelectronics, Chinese Academy of Sciences report a hybrid flash-SRAM CIM design to support on-chip plastic learning in a 14 nm FinFET process[13].

Since Tu *et al*. from Tsinghua University proposed the first FP CIM on ISSCC in 2022[14], the CIM community has been actively studying this new direction to meet the growing high-precision demand of both AI inference and training, especially in this generative AI era. Wang *et al*. from Tsinghua University propose an FP SRAM-CIM macro based on an emerging dynamic POSIT8 data format, which utilizes the low bitwidth to achieve comparable accuracy to the BF16 format[15]. Wen *et al*. from National Tsing Hua University present an FP ReRAM-CIM macro, with a kernel-wise weight pre-alignment scheme and a rescheduled multi-bit input compression scheme to suppress the amount of truncated data by 1.96–2.47× and reduce MAC operation cycles by 4.73×[16].

Moreover, researchers are collaborating and considering challenges from a systematic perspective to help CIM integrate into real computing systems. A DCIM-based neural visual-enhancement engine (NVE) is fabricated in the 3 nm process through the collaboration of MediaTek and TSMC[17]. Wang *et al*. from the University of Texas at Austin present Vecim, a RISC-V vector co-processor integrated with a CIM-based vector register file, using foundry SRAM cells in 65 nm CMOS for efficient high-performance computing[18]. On the other hand, the fusion with RISC-V also enhances the programmability of CIM-based systems.

Trend 3: DSA for embedded vision processors

When we move towards more autonomous and collaborative AI computing, visual perception of physical environments becomes a fundamental capability for future integrated smart systems. In 2023, Apple Vision Pro[19] reinvigorated the landscape of augmented reality (AR) and virtual reality (VR), blending digital elements with our physical surroundings (a.k.a. spatial computing). Moreover, emerging embodied AI hinges on agile visual perception and LLM-embedded processing systems, enabling a deeper understanding of our world (e.g., Figure 01 robot[20]). However, these vision workloads need to directly communicate with humans and give fast feedback during runtime. A slow response may significantly hurt user experience, especially in autonomous driving tasks. Also, edge devices usually have strict constraints on device weight, leading to a layout with limited batteries. To drive these devices continuously, ultra-low power consumption is required for the underlying hardware. As more real-time and resource-constrained edge systems are supporting intelligent vision tasks, we observed a trend to design dedicated embedded vision processors.

ISSCC 2024 sheds light on many research prototypes and commercial products on this trend. Park *et al*. from KAIST integrate neural radiance fields (NeRF), simultaneous localization and mapping (SLAM), and sparse mixture-of-expert (SMoE) in their space-mate[21], a fast and low-power NeRF-SLAM edge accelerator, which solves irregular SMoE expert loading patterns by out-of-order MoE router, large mapping energy expense by familiar region pruning, and dual-mode sparsity

by heterogeneous coarse-grained sparse core. Ryu *et al*., also from KAIST, design an accelerator for vanilla-NeRF-based instant 3D modeling and real-time rendering, called NeuGPU[22]. They use segmented hashing tables with data tiling to reduce on-chip storage pressure and attention-based hybrid interpolation units to alleviate bank conflict costs. Also, by their exploration of similar activation characteristics in NeRF, they compressed similar values in feature vectors of adjacent samples. Nose *et al*. from Renesas Electronics propose a heterogenous processor for multi-task and real-time robot applications[23], combining vision recognition with planning and control based on the cooperation among a dynamically reconfigurable processor, AI accelerator, and embedded CPU.

Trend 4: DSA for solver accelerators

The optimization problem solver is another important component of integrated smart systems, which can take the role of intelligent decision-making and planning. Optimization problems across various domains, including modeling, controlling, and scheduling, are addressed by corresponding solver algorithms. However, given that the large and complex solution space should be efficiently explored in real-time systems, hardware implementations of solver algorithms face strict requirements in terms of low latency, high accuracy, and high robustness.

In ISSCC 2024, we observe an emerging trend of DSA for solver accelerators. One of the popular solvers is the Ising machine solver for combinatorial optimization problems (COP). The Ising machine's significance lies in its ability to solve COPs with the nondeterministic polynomial-time (NP) complexity costing only polynomial overhead. Together with the hardware-friendly dataflow, Ising machine accelerators achieve significant speedup and efficiency improvement for solving COPs. Chu *et al*. from National Taiwan University propose an annealing-based Ising machine processor for large-scale autonomous navigation with integrated mapping workflow[24]. Song *et al*. from Peking University design an eDRAM-based continuous-time Ising machine with embedded annealing and leaked negative feedback[25]. Bae *et al*. from the UCSB propose two chips for scalable SRAM-based Ising macro with enhanced chimera topology[26] and continuous-time latch-based Ising computer using massively parallel random-number generations and replica equalizations[27].

Apart from the Ising machine, there are also accelerators for other solver algorithms in ISSCC 2024. Shim *et al*. from UCSB propose VIP-SAT to solve the Boolean satisfiability (SAT) problem with a scalable digital in-memory computing dataflow and hardware-software co-design method[28]. Ju *et al*. from Northwestern University accelerate real-time partial differential equation (PDE) solver on edge devices by reconfiguring its architecture between advanced physics-informed neural network (PINN) mode for low latency and traditional finite element method (FEM) mode for high accuracy[29].

In conclusion, we discuss four exciting ML accelerator research trends in ISSCC 2024: **ML chips for generative AI, CIM innovation from circuits to systems, DSA for embedded vision processors, and DSA for solver accelerators**. With the goal of efficient generative AI and beyond-AI computing, we believe future ML accelerators will realize general intelligence in diverse application scenarios.

Acknowledgments

This research was supported in part by ACCESS - AI Chip Center for Emerging Smart Systems, sponsored by InnoHK funding, Hong Kong SAR, and HKUST-HKUST(GZ) 20 for 20 Cross-campus Collaborative Research Scheme C031.

References

[1] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint, 2021

[2] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv preprint, 2023

[3] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. International Conference on Machine Learning (ICML), 2021

[4] Mu C, Zheng J P, Chen C X. Beyond convolutional neural networks computing: New trends on ISSCC 2023 machine learning chips. J Semicond, 2023, 44, 050203

[5] Alben J. Computing in the era of generative AI. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 26

[6] Smith A, Chapman E, Patel C, et al. AMD InstinctTM MI300 series modular chiplet package–HPC and AI accelerator for exa-class systems. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 490

[7] Guo R Q, Wang L, Chen X F, et al. A 28nm 74.34TFLOPS/W BF16 heterogenous CIM-based accelerator exploiting denoising-similarity for diffusion models. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 362

[8] Kim S, Kim S, Jo W, et al. C-transformer: A 2.6-18.1μJ/token homogeneous DNN-transformer/spiking-transformer processor with big-little network and implicit weight generation for large language models. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 368

[9] ISSCC 2024 forum 2: Energy-efficient AI-computing systems for large-language models. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 593

[10] Fujiwara H, Mori H, Zhao W C, et al. A 3nm, 32.5TOPS/W, 55.0TOPS/mm^2 and 3.78Mb/mm^2 fully-digital compute-in-memory macro supporting INT12 × INT12 with a parallel-MAC architecture and foundry 6T-SRAM bit cell. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 572

[11] He Y F, Fan S P, Li X, et al. A 28nm 2.4Mb/mm^2 6.9-16.3TOPS/mm^2 eDRAM-LUT-based digital-computing-in-memory macro with in-memory encoding and refreshing. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 578

[12] Guo A, Chen X, Dong F Y, et al. A 22nm 64kb lightning-like hybrid computing-in-memory macro with a compressed adder tree and analog-storage quantizers for transformer and CNNs. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 570

[13] Wang L F, Li W Z, Zhou Z D, et al. A flash-SRAM-ADC-fused plastic computing-in-memory macro for learning in neural networks in a standard 14nm FinFET process. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 582

[14] Tu F B, Wang Y Q, Wu Z H, et al. A 28nm 29.2TFLOPS/W BF16 and 36.5TOPS/W INT8 reconfigurable digital CIM processor with unified FP/INT pipeline and bitwise In-memory booth multiplication for cloud deep learning acceleration. 2022 IEEE International Solid-State Circuits Conference (ISSCC), 2022, 1

[15] Wang Y, Yang X L, Qin Y B, et al. A 28nm 83.23TFLOPS/W POSIT-based compute-in-memory macro for high-accuracy AI applications. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 566

[16] Wen T H, Hsu H H, Khwa W S, et al. A 22nm 16Mb floating-point

ReRAM compute-in-memory macro with 31.2TFLOPS/W for AI edge devices. 2024 IEEE International Solid State Circuits Conference (ISSCC), 2024

[17] Shih M E, Hsieh S W, Tsai P Y, et al. NVE: A 3nm 23.2TOPS/W 12b-digital-CIM-based neural engine for high-resolution visual-quality enhancement on smart devices. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 360

[18] Wang Y P, Yang M T, Lo C P, et al. Vecim: A 289.13GOPS/W RISC-V vector co-processor with compute-in-memory vector register file for efficient high-performance computing. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 492

[19] Apple Vision Pro, https://www.apple.com/apple-vision-pro/

[20] Figure 01 robot, https://www.figure.ai/

[21] Park G, Song S, Sang H Y, et al. Space-mate: A 303.5mW real-time sparse mixture-of-experts-based NeRF-SLAM processor for mobile spatial computing. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 374

[22] Ryu J, Kwon H, Park W, et al. NeuGPU: A 18.5mJ/iter neural-graphics processing unit for instant-modeling and real-time rendering with segmented-hashing architecture. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 372

[23] Nose K, Fujii T, Togawa K, et al. A 23.9TOPS/W @ 0.8V, 130TOPS AI accelerator with 16 × performance-accelerable pruning in 14nm heterogeneous embedded MPU for real-time robot applications. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 364

[24] Chu Y C, Lin Y C, Lo Y C, et al. A fully integrated annealing processor for large-scale autonomous navigation optimization. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 488

[25] Song J H, Wu Z H, Tang X Y, et al. A variation-tolerant In-eDRAM continuous-time Ising machine featuring 15-level coefficients and leaked negative-feedback annealing. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 490

[26] Bae J, Shim C, Kim B. E-chimera: A scalable SRAM-based Ising macro with enhanced-chimera topology for solving combinatorial optimization problems within memory. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 286

[27] Bae J, Koo J, Shim C, et al. LISA: A 576 × 4 all-in-one replica-spins continuous-time latch-based Ising computer using massively-parallel random-number generations and replica equalizations. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 284

[28] Shim C, Bae J, Kim B. VIP-sat: A Boolean satisfiability solver featuring 5 × 12 variable in-memory processing elements with 98% solvability for 50-variables 218-clauses 3-SAT problems. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 486

[29] Ju Y H, Xu G Q, Gu J. A 28nm physics computing unit supporting emerging physics-informed neural network and finite element method for real-time scientific computing on edge devices. 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, 366

Bohan Yang is currently a senior undergraduate at the School of the Gifted Young, University of Science and Technology of China, Hefei, China. He is also a visiting intern at the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. His research interests include modern computer architecture, accelerators for emerging workloads, and software hardware co-design.

Jia Chen is currently a Postdoc researcher at the AI Chip Center for Emerging Smart Systems (ACCESS), The Hong Kong University of Science and Technology. She received her Ph.D. degree in microelectronics and solid-state electronics from Huazhong University of Science and Technology in 2021. Her research interests include emerging non-volatile memory devices and computing-in-memory related circuit design.

Fengbin Tu received the B.S. degree from the School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China, in 2013, and received the Ph.D. degree from the Institute of Microelectronics, Tsinghua University, Beijing, China, in 2019. Dr. Tu is currently an Assistant Professor at the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. He was a Postdoctoral Fellow at the AI Chip Center for Emerging Smart Systems (ACCESS), Hong Kong, China, from 2022 to 2023, and a Postdoctoral Scholar at the Scalable Energy-efficient Architecture Lab (SEAL), the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA, from 2019 to 2022. His research interests include AI chip, computer architecture, reconfigurable computing, and computing-in-memory. His AI chips ReDCIM and Thinker won the 2023 Top-10 Research Advances in China Semiconductors and 2017 ISLPED Design Contest Award, respectively. He has published two books, *Architecture Design and Memory Optimization for Neural Network Accelerators* and *Artificial Intelligence Chip Design*. His research works appeared at top conferences and journals on integrated circuits and computer architecture, including ISSCC, JSSC, DAC, ISCA, and MICRO.