MEGATRON: a 28nm Analog PCM CiM/Digital System-on-Chip for Edge GenAI at 57.5 TOPS/W and 1.52 Mparam/mm^2

Alessandro Nadalini*†, Angelo Garofalo*†, Lorenzo Greco*†, Andrea Belano†‡, Alessio Antolini†, Francesco Zavalloni†, Andrea Lico‡, Riccardo Zurla§, Emanuela Calvetti§, Luigi Croce§, Marco Pasotti§, Alessandro Cabrini¶, Eleonora Franchi Scarselli†, Davide Rossi†‡, Francesco Conti†‡
†University of Bologna, Italy, ‡Chips-IT, Italy, §STMicroelectronics, Italy, ¶University of Pavia, Italy, *Equal contribution

***Abstract*—We present MEGATRON, a heterogeneous Edge GenAI System-on-Chip in 28nm FD-SOI CMOS technology combining a non-volatile analog in-memory-computing engine based on a 4Mi-cell phase-change memory (PCM) array with a digital RISC-V-based flexible neural processing unit. MEGATRON demonstrates up to 3.5 TOPS/W using the RISC-V processors and 57.5 TOPS/W with PCiM, at a storage density of 1.52 Mparam/mm^2 with 4-bit effective weight precision.**

I. INTRODUCTION

The recent, exponential growth in capabilities of generative artificial intelligence (GenAI) algorithms such as Large Language Models (LLMs) and Mamba/SSMs is bringing about a revolution in the capabilities of wearables, robots, and other autonomous devices. Yet, the dependence on computation performed in remote datacenters raises many challenges in terms of latency, availability, and privacy. Edge GenAI aims at moving at least part of the burden of execution from datacenters to the edge devices themselves, similarly to what was done in the past years with perception AI algorithms such as CNNs – however, Edge GenAI algorithms, such as Small Language Models (SLMs), have much more challenging characteristics in terms of memory footprint (10-100× more parameters; need for space for key-value (KV) caching) and computational patterns (e.g., the decode phase of LLMs is typically memory-bound due to the combined effect of low weight reuse in matrix-vector products and KV caching).

Non-volatile analog in-memory-computing (NV-AIMC) based on resistive memories (ReRAMs) [1] and phase-change memories (PCMs) [2] [3] is attractive for Edge GenAI: putting high density non-volatile parameter storage with computation effectively removes all memory traffic due to network weights, leaving only that of KV caching. However, it also raises many challenges for system-level integration and scaling to advanced technologies. Computational PCMs have recently attracted attention due to the possibility to achieve high weight density and circumvent data drift. However, previous works have focused primarily on technological aspects while neglecting system-level impacts of the introduction of this technology.

To fill in this gap, we introduce MEGATRON, a prototype heterogeneous System-on-Chip in 28nm FD-SOI CMOS STMicroelectronics technology for Edge GenAI combining

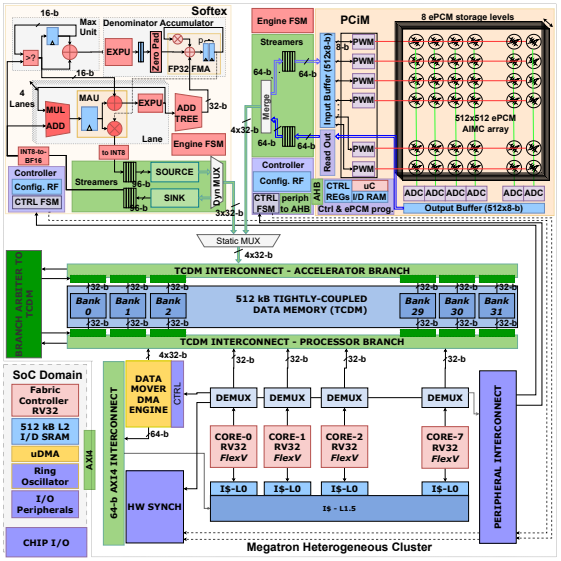

Fig. 1. MEGATRON SoC architecture and Heterogeneous Cluster domain.

NV-AIMC based on a 4Mi-cells phase-change memories (PCMs) with a fully flexible and programmable digital neural processing unit (NPU) based on a RISC-V DSP cluster. Our contributions are the following:

1) we show industry-leading on-chip parameter density in 28nm node of 1.52 Mparam/mm^2;
2) we propose a mixed analog PCM/digital system, achieving up to 57.5 TOPS/W on linear layers at system-level with 8-bit activations and ENOB=4 weights;
3) we extrapolate end-to-end fully-on-chip execution of one layer of an emerging SLM (SmollLM-135M).

II. SOC ARCHITECTURE

MEGATRON includes two domains: *System-on-Chip (SoC)* and *Heterogeneous Cluster (HC)*, as shown in Figure 1. The SoC domain acts as an on-chip RISC-V `RV32IMC` microcontroller for testing purposes, including I/O peripherals, an autonomous I/O uDMA, 512 KiB of L2 SRAM for data and instructions, and a ring oscillator for clock generation. The

HC domain comprises the primary contribution of our work, targeting on-device perceptive and generative Transformer models. It includes eight RISC-V `RV32IMCFXpulpNN` cores (**FlexV**) [4], a DMA engine, a Phase-Change Compute-in-Memory accelerator (**PCiM**) and a digital accelerator for non-linear operators such as Softmax and GeLU (**Softex**). All units are connected to a 32-bank, 512 KiB Tightly-Coupled Data Memory (TCDM) accessible through a low-latency (1-cycle) hybrid interconnect enabling simultaneous access from an accelerator branch (from PCiM, Softex) and a processor branch (from all FlexVs and the cluster DMA), forming a single tightly-coupled unit with a total of 512b/cycle of aggregate bandwidth (30.72 GB/s @ 480 MHz). Both Softex and PCiM access the TCDM accelerator branch through a set of *streamers* that expose inputs and outputs to their respective microarchitecture as ready-valid data streams. Both accelerators are controlled via a memory-mapped register interface; the access of Softex and PCiM is alternate and regulated by a static multiplexer.

The computational core of PCiM is a phase-change memory (PCM) analog in-memory computing (AIMC) macro [5] in STMicroelectronics 28nm FDSOI CMOS, wrapped within a digital interface. It accelerates products between a stationary (non-volatile) matrix and a streamer-produced vector, exploiting analog storage and computation within the PCM array itself. PCiM stores 2M signed 4-bit weights (ENOB) across eight alternative *scenarios*, supporting 512×512 matrix-vector products with signed 8-bit inputs and up to 11-bit output precision. Each weight employs two PCM cells to encode positive and negative values. During AIMC computation, inputs are encoded as time intervals on PCM Wordlines (WLs), while dedicated Bitline (BL) voltage regulators overcome PCM current-voltage non-linearity. BL currents are integrated over a time window and digitized via current-controlled oscillator ADCs. An internal 32-bit microcontroller handles configuration, weight programming, and calibration, while runtime execution uses a digital FSM interface. Double buffers on decoupled input/output channels enable pipelined dataflow with the streamers, supporting both GEMM and MVM kernels for the *prefill* and *decode* of autoregressive Transformers.

Softex is a hardware accelerator for non-linearities such as Softmax and GeLU, integrated as a cooperative accelerator sharing the TCDM with the FlexV cores and with PCiM. Its datapath features 4 lanes, each comprising a BF16 Multiplication-and-Addition Unit (MAU) and an Exponential Unit (EXPU) implementing a hardware-friendly approximation of the exponential function based on an extension of Schraudolph's method with polynomial mantissa correction. Softmax computation is organized in three phases: accumulation (max search and denominator calculation), inversion (Newton-Raphson reciprocal), and normalization (output probability computation). Softex communicates with the rest of the HC through a ready/valid streaming interface connected to the TCDM accelerator branch: configurable INT8-to-BF16 and BF16-to-INT8 data converters are used for compatibility with the 8-bit activation format of PCiM.

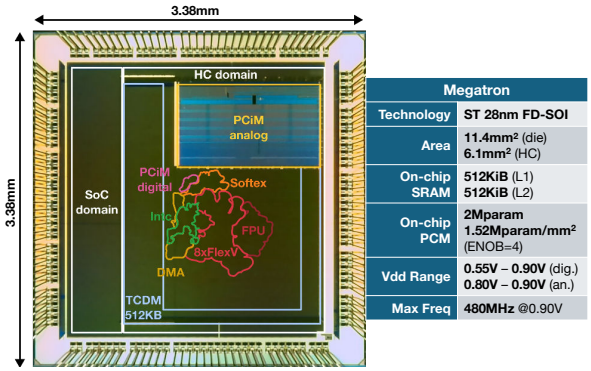

Fig. 2. MEGATRON chip microphotograph, highlighting the SoC and HC domains and the main computing blocks within the HC.

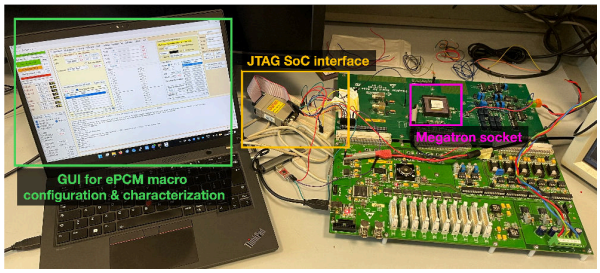

Fig. 3. Experimental setup for MEGATRON test chip characterization.

III. MEASUREMENTS AND BENCHMARKING

Figure 2 shows a microphotograph of the MEGATRON chip, fabricated in 28nm FD-SOI CMOS STMicroelectronics technology. The die size is 3.38mm×3.38mm. We focused our analysis on the HC domain (6.1mm^2), which includes the PCiM macro (1.34 mm^2, yielding a storage density of 1.49 Mparam/mm^2). The same supply powers both the digital and analog IPs of MEGATRON. The chip is working between 0.55 V–0.90 V for what concerns the digital components and SRAM memories, and between 0.80 V–0.90 V for the analog PCiM. We characterized the HC digital part across operating voltages, frequencies, and workloads at room temperature (25 °C). As shown in Figure 3, the MEGATRON prototype is socketed on a dedicated carrier board exposing probe points for oscilloscope measurements and current monitoring via digital multimeter. The carrier board mounts on a motherboard hosting an FPGA for PCM macro characterization from a laptop, with digital architecture access via a JTAG interface connected to the SoC debug module.

Figure 4a reports experimental results in terms of frequency and HC domain power across the 0.55 V–0.90 V supply voltage range, targeting a matrix multiplication kernel executed by the FlexV cores under several bit-width configurations and the operation of Softex. MEGATRON can operate at a maximum frequency of 480 MHz@0.90 V in the highest performance operating point, and at up to 100 MHz@0.55 V in the slowest operating point. In Figure 4b we characterize MEGATRON in terms of energy efficiency, considering the operation of

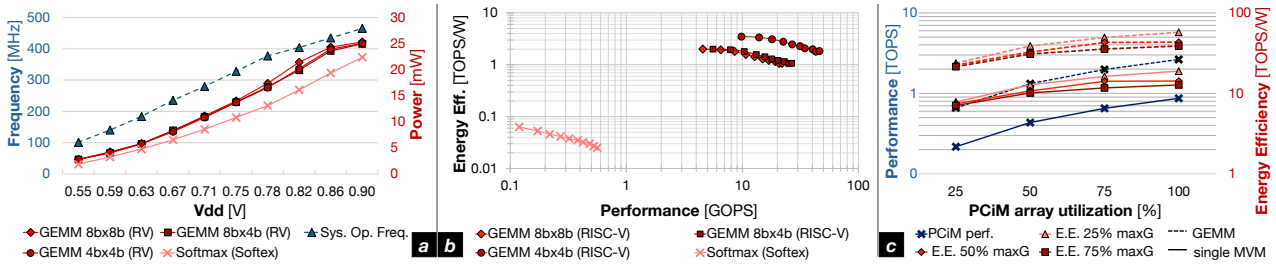

Fig. 4. *a)* Voltage/frequency/power sweep on MEGATRON prototype; *b)* Performance and energy efficiency of digital blocks; *c)* Performance and energy efficiency of PCiM sweeping conductance (25–50–75% of the maximum G), PCiM array utilization and kernel (GEMM vs single MVM). For the Softmax kernel we adopt the convention of 1 OP/token.

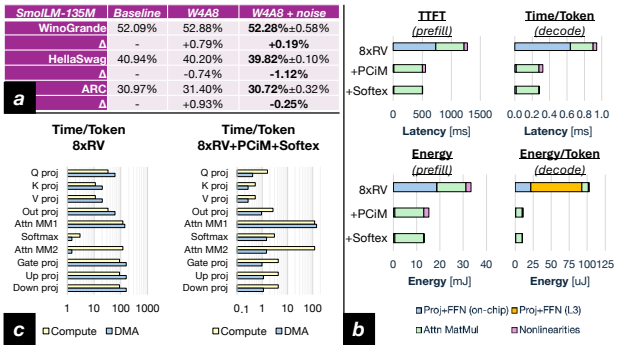

Fig. 5. Evaluation of 1 layer of *SmolLM-135M* with sequence size $S = 2048$ on MEGATRON-XL: *a)* accuracy on WinoGrande, HellaSwag and ARC-Challenge under full-precision, quantized (*W4A8*), and analog (*W4A8 + noise*) conditions; *b)* latency and energy analysis in prefill and decode phase with/without on-chip PCiM and Softex; *c)* detail of decode phase time-per-token for one SmolLM-135M layer.

digital IPs. FlexV achieves up to 3.5 TOPS/W in the best-efficiency operating point (0.55V 4×4-bit), providing a fully programmable NPU with efficiency comparable to custom datapath; in the best-performance point (0.90V), it achieves 46 GOPS at 1.84 TOPS/W. Softex can process up to 560 Mtoken/s at 25 Gtoken/s/W at 0.90V, or up to 63 Gtoken/s/W at 0.55V. Figure 4c analyzes the performance and energy efficiency of the PCiM array while sweeping average conductance, PCM utilization and target kernel (GEMM vs MVM). PCiM achieves up to 868 GOPS @ 18.9 TOPS/W in MVM and 2.64 TOPS @ 57.5 TOPS/W in GEMM.

IV. EVALUATION ON SMOLLM-135M

To contextualize the PCM/digital integration technique and MEGATRON capabilities, we evaluated *SmolLM-135M* by HuggingFace as a representative Edge GenAI workload. Since PCM is non-volatile, the largest on-chip deployable network is capped by storage capacity (2Mparam for MEGATRON). We therefore modeled MEGATRON-XL, a larger configuration hosting two multiplexed PCiM macros to fit one SmolLM-135M layer (3.54×10^6 parameters) and a larger 8MiB on-chip L2 SRAM for residuals and KV-cache. Data transfers were modeled both on-chip (L2–L1) and off-chip (L3–L1, for parameters without PCiM), using measured on-chip performance and power, and an LPDDR6 device [6] at 60 GB/s and 5 pJ/bit for off-chip access. Weights were 4-bit quantized with PCiM-characterized noise, and network accuracy was evaluated in PyTorch using `transformers` and `brevitas`, with analog noise modeled following Antolini et al. [7].

Figure 5 reports accuracy, performance (time-to-first-token and time/token), and energy for SmolLM-135M. First, Fig. 5a confirms that quantization and analog noise do not significantly degrade accuracy on three representative benchmarks. In prefill (Fig. 5b left), PCiM yields $\sim 60\times$ speedup on static-weight layers; combined with Softex ($\sim 14\times$ on non-linearities), this gives $\sim 2.5\times$ overall per-layer speedup, with similar gains for what concerns energy; L3 access in the absence of PCM has limited impact due to on-chip data reuse in compute-bound conditions. In decode (Fig. 5b right), gains are even larger: decode is typically memory-bound, but PCiM eliminates L3 weight loading by keeping weights on-chip, restoring compute-bound operation. This yields $\sim 3.3\times$ time/token improvement over the baseline RISC-V cluster, detailed per-operator in Fig. 5c. The energy benefit is particularly striking in memory-bound conditions due to eliminated L3 traffic, achieving $\sim 10.5\times$ better energy efficiency overall.

V. COMPARISON WITH STATE-OF-THE-ART

Fig. 6 compares MEGATRON against state-of-the-art edge-AI systems. Considering non-volatile ReRAM-based [1] and PCM-based [2] CiM-only SoC, we achieve 2.44× and 27.4× higher CiM storage density, respectively, thanks to our multi-level PCiM array, at comparable bit-equivalent energy efficiency, noting that their figures benefit from 55% input sparsity, which we do not assume. Against *Hermes* [3], which integrates 64 PCM-based ACiM tiles and a digital LSTM accelerator, we achieve comparable single-tile performance, while achieving overall comparable area efficiency, 5.5× higher energy efficiency, and 32.3× greater CiM storage density. Compared to heterogeneous SoCs with SRAM-based [8] and MRAM-based [9] ACiM, we outperform both in area efficiency (2.98× and 15.9×, respectively) and CiM storage density (23.6× and 17.5×). While *Diana* [8] achieves higher peak energy efficiency, its SRAM-based ACiM lacks non-volatile parameter storage capabilities, potentially incurring higher end-to-end energy costs from memory-hierarchy

	ISSCC26 [1]	ISSCC22 [2]	*Hermes* Nature23 [3]	*Diana* ISSCC22 [8]	*Falcon* ESSERC25 [9]	ISSCC 23 [10]	ISSCC 26 [11]	*Siracusa* JSSCC 24 [12]	This Work
Type	ACIM	ACIM	Custom Acc. w/ ACIM	Het. SoC w/ ACIM	Het. SoC w/ ACIM and NMC	Het. SoC w/ DCIM	Het. SoC w/ NMC	Het. SoC w/ NMC	Het. SoC w/ AIMC
Application	Mamba/Transformers/CNN	Edge Tiny-AI	CNN/LSTM	AI-IoT	CNN + Attention in Obj. Detection	CNN/LSTM/RNN	LLMs	Edge AI - XR	Edge GenAI
Technology	22nm	40nm CMOS	14nm	GF 22nm	28nm	ST 18nm FD-SOI	55nm	16nm FinFet	ST 28nm FD-SOI
CIM Technology	1T1R ReRAM	1T1R PCM	8T4R PCM	Analog SRAM	MRAM	Digital SRAM	None	None	2T2R ePCM w/ 8 storage levels
Die Area [mm2]	38.55	18	24 (1 PCM tile: 1.39 mm2)	10.24; 2.29 (AIMC core)	12.96	4.2 (IMNPU cluster)	55.98 (digital die) 22.98 (stacked ReRAM)	10.7 (cluster)	1.38 (AIMC) 7.5 (cluster) - 14.1 (SoC)
Supply Voltage [V]	0.65 -0.8	0.75 - 0.9	0.8	0.45 - 0.9	0.7 - 1.15; 1.8 (MRAM)	0.525 - 1	0.89 - 1.04 (logic die)	0.65 - 0.8	0.55V - 0.9V
Power [mW]	N.A.	N.A.	N.A.	10 -129 (digital)	29.5 - 196.5	N.A.	265 - 3060	151 - 332	2 - 25 (digital); 68 (AIMC)
CIM storage capacity	96 Mb	2M params (8x256Rx1024C)	4.192 Mparams (64 256x256 arrays)	589.8k weights (*)	4.5 Mb	2 Mb	N.A.	N.A.	2M params (512x512 w/ 8 levels)
CIM Storage Density	623.4 kparams/mm2 (*)(**)(NV)	55.6 kparams/mm2 (*)(**)(NV)	47.1 kparams/mm2 (*)(✣)(NV)	64.39 kparams/mm2 (*)(**)(✣)(V)	86.8 kparams/mm2 (*)(**)(✳)(NV)	124.8 kparams/mm2 (*)(**)(✳)(V)	N.A.	N.A.	**1.52 Mparams/mm2 (**)(NV)**
Best Perf. [TOPS]	N.A.	4bIN-4bAW-11bOUT: 0.829 8bIN-8bAW-19bOUT: 0.475 (55% sparsity)	8bIN-4bAW-8bOUT: 63.1 (1 PCM Tile: 1.008)	7bIN-TAW-7bOUT: 29.5 (ACIM) 2bIN-2bW-8bOUT: 0.18 (digital)	1bIN-1bW-3bOUT: 49.15 (ACIM) System (CNN + FP8 att.): 1.928	1bIN-1bW-16bOUT: 229 4bIN-4bW-16bOUT: 57	8bIN-4bW/8bIN-8bW-8bOUT: 2.33	8bIN-2bW-8bOUT: 1.95 (acc.) 8bIN-8bW-32bOUT: 0.03 (RV)	8bIN-4bAW-8bOUT: 2.6 (AIMC) 8bIN-8bW-32bOUT: 0.02 (RV)
Best En. Eff. [TOPS/W]	8bIN-8bAW: 112.8 - 170.2 (55% - 90% sparsity)	4bIN-4bAW-11bOUT: 57.6 - 182 8bIN-8bAW-19bOUT: 205 - 65 (55% - 90% sparsity)	8bIN-4bAW-8bOUT: 9.76 (1 PCM tile: 10.5)	7bIN-TAW-7bOUT: 600 (ACIM) 2bIN-2bW-8bOUT: 4.1 (digital)	1bIN-1bW-3bOUT: 750.18 (ACIM) System (CNN + FP8 att.): 16.34	1bIN-1bW-16bOUT: 310 4bIN-4bW-16bOUT: 77	8bIN-4bW/8bIN-8bW-8bOUT: 1.92 (*)	8bIN-2bW-8bOUT: 8.84 (acc.) 8bIN-8bW-32bOUT: 2.68 (RV)	8bIN-4bAW-8bOUT: 57.5 (AIMC) 8bIN-8bW-32bOUT: 1.64 (RV)
Best Area Eff. [TOPS/mm2]	N.A.	4bIN-4bAW-11bOUT: 0.05 8bIN-8bAW-19bOUT: 0.03 (55% sparsity)	8bIN-4bAW-8bOUT: 1.55 (1 PCM tile: 1.59)	7bIN-TAW-7bOUT: 2.88 (ACIM) 2bIN-2bW-8bOUT: 0.021 (digital)	1bIN-1bW-3bOUT: 3.79 (ACIM) System (CNN + FP8 att.): 0.15	1bIN-1bW-16bOUT: 54 4bIN-4bW-16bOUT: 13.6	8bIN-4bW/8bIN-8bW-8bOUT: 0.04 (*)	8bIN-2bW-8bOUT: 0.18 (acc.) 8bIN-8bW-32bOUT: 0.003 (RV)	8bIN-4bAW-8bOUT: 1.88 (AIMC) 8bIN-8bW-32bOUT: 0.003 (RV)
Peak bin. Equiv. Perf. [TOPS] (***)	N.A.	7.8 (55% sparsity)	2019 (1 PCM tile: 32.3)	206.5 (ACIM) 1.26 (digital)	49.15 (ACIM)	229	74.56	31.2 (N-EUREKA)	83.2 (AIMC) 1.41 (RV)
Peak bin. Equiv. En. Eff. [TOPS/W] (***)	7219 (55% sparsity)	1436 (55% sparsity)	312.32 (1 PCM tile: 336)	**4200 (ACIM)** 0.15 (digital)	750.18 (ACIM)	310	61.44	141.44 (N-EUREKA)	**1840** (AIMC) 105 (RV)
Peak bin. Equiv. Area Eff. [TOPS/mm2] (***)	N.A.	0.44 (55% sparsity)	49.6 (1 PCM tile: 50.88)	20.16 (ACIM) 0.15 (digital)	3.79 (ACIM)	54	1.28	2.912 (N-EUREKA)	**60.16** (AIMC) 0.19 (RV)

(V) = volatile; (NV): non-volatile; TAW = ternary analog weights; AW = analog weights; (*) derived from provided data; (**) 4-b equivalent (ENOB=4); (***) metrics scaled to 1bIN-1bW/AW; (✣) derived considering ACIM area only; (✳) Not possible to evaluate ACIM only area from the data in the paper.

Fig. 6. Comparison with state-of-the-art edge AI systems.

parameter movement. Against a SRAM-based digital CiM SoC [10], despite its higher throughput due to the integration of 64 32×1024-b DCiM tiles, we achieve 1.2×, 5.94× higher area and energy efficiency, respectively, with 12.2× greater storage density (non-volatile, in our case). Finally, compared to traditional near-memory edge-ai computing SoCs [11], [12], we deliver 30× and 13× higher energy-efficiency, and 47× and 21× higher area-efficiency, respectively, at comparable or better performance, though their fully digital accelerators retain the advantage of iso-accuracy AI execution.

VI. Conclusion

The MEGATRON test chip demonstrates heterogeneous analog PCM/digital integration for Edge GenAI in 28nm FD-SOI, combining a 4Mi-cell, 2Mparam PCM compute-in-memory accelerator, RISC-V DSP cluster, and Softex non-linear accelerator to achieve 57.5 TOPS/W at 1.52 Mparam/mm^2. SmolLM-135M benchmarking confirms PCM-based in-memory computing reduces latency and energy-especially in the memory-bound decode phase-with negligible accuracy loss, establishing NV-AIMC as a viable path to fully on-chip Small Language Model execution at the edge.

References

[1] H.-H. Hsu *et al.*, "A 22nm 96Mb 50.6-to-90.2TFLOPS/W Non-Linear MLC ReRAM CIM Macro with High-Retention for Mamba/Transformer/CNN," in *2026 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 69, pp. 516–518, 2026.

[2] W.-S. Khwa *et al.*, "A 40-nm, 2M-Cell, 8b-Precision, Hybrid SLC-MLC PCM Computing-in-Memory Macro with 20.5 - 65.0TOPS/W for Tiny-Al Edge Devices," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, pp. 1–3, 2022.

[3] M. Le Gallo *et al.*, "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," *Nature Electronics*, vol. 6, no. 9, pp. 680–693, 2023.

[4] A. Nadalini *et al.*, "A 3 TOPS/W RISC-V parallel cluster for inference of fine-grain mixed-precision quantized neural networks," in *2023 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 1–6, IEEE, 2023.

[5] M. Pasotti *et al.*, "28 M Weights × TOPs /W/mm^2 PCM-Based Analog in-Memory Computing Core with 8 512 × 512-Weight Layers in 28 nm FD-SOI CMOS," in *2025 IEEE European Solid-State Electronics Research Conference (ESSERC)*, pp. 129–132, 2025.

[6] D. Christ *et al.*, "Performance and power analysis of lpddr6," in *2025 Cross-Disciplinary Conference on Memory-Centric Computing (CCMCC)*, pp. 1–7, 2025.

[7] A. Antolini *et al.*, "Controlled Acceleration of PCM Cells Time Drift Through On-Chip Current-Induced Annealing for AIMC Multilevel MVM Computation," *IEEE Transactions on Electron Devices*, vol. 72, no. 1, pp. 215–221, 2025.

[8] K. Ueyoshi *et al.*, "DIANA: An End-to-End Energy-Efficient Digital and ANAlog Hybrid Neural Network SoC," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, pp. 1–3, 2022.

[9] W. Xie *et al.*, "Falcon: An Event-Driven Object Tracking SoC with Activation-Skip and Weight-Compression MRAM PIM and Heterogenous Q/K/V Read-Only-Once Attention Flow," in *2025 IEEE European Solid-State Electronics Research Conference (ESSERC)*, pp. 125–128, 2025.

[10] G. Desoli *et al.*, "16.7 A 40-310TOPS/W SRAM-Based All-Digital Up to 4b In-Memory Computing Multi-Tiled NN Accelerator in FD-SOI 18nm for Deep-Learning Edge Applications," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 260–262, 2023.

[11] P. Dong *et al.*, "A 14.08-to-135.69Token/s ReRAM-on-Logic Stacked Outlier-Free Large-Language-Model Accelerator with Block-Clustered Weight-Compression and Adaptive Parallel-Speculative-Decoding," in *2026 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 69, pp. 532–534, 2026.

[12] A. S. Prasad *et al.*, "Siracusa: A 16 nm Heterogenous RISC-V SoC for Extended Reality With At-MRAM Neural Engine," *IEEE Journal of Solid-State Circuits*, vol. 59, no. 7, pp. 2055–2069, 2024.