CHIMERA: A Flexible and Scalable 3.1 TOPS/W AI-MCU with Transformer Accelerator and 563 Gb/s Shared-L2 Memory Subsystem with QoS Guarantees

Lorenzo Leone*, Philip Wiese*, Gamze Islamoglu*, Michael Rogenmoser*
Davide Rossi†‡, Francesco Conti†, Luca Benini*†
*ETH Zürich, Zürich, Switzerland, †Università di Bologna, Bologna, Italy, ‡Chips-IT, Pavia, Italy
*{lleone, wiesep, gislamoglu, michaero, lbenini}@iis.ee.ethz.ch, †{davide.rossi, f.conti}@unibo.it

***Abstract*—We present Chimera, a flexible and scalable Microcontroller Unit (MCU) designed to accelerate real-time inference of rapidly evolving transformer-based models at the ultra-low-power edge (hundred of mW). The chip, implemented in 22 nm FDX technology, integrates a transformer accelerator tightly coupled within a compute cluster featuring nine general-purpose RV32IMA cores. Scalability extends to the memory hierarchy through a novel L2 memory island subsystem, which enables data sharing across multiple clusters while delivering 563 Gb/s aggregate bandwidth. The L2 subsystem enforces quality-of-service guarantees for latency-critical traffic, achieving up to 16× latency reduction. Chimera achieves peak energy and area efficiencies of 3.1 TOPS/W and 281 GOPS/mm², demonstrating 1.37× higher energy efficiency and up to 100× higher area efficiency compared to State of the Art (SoA) SoCs. Compared to SoA standalone accelerators, Chimera achieves comparable energy efficiency and up to 1.8× higher area efficiency.**

***Index Terms*—Transformers, Edge-AI, MCU, QoS, Attention**

I. Introduction

The increasing diffusion of Artificial Intelligence (AI) workloads, such as Natural Language Processing (NLP) and speech recognition, in edge and TinyML systems drives the need for high-throughput AI-accelerated Microcontroller Units (AI-MCUs) supporting real-time-constrained execution under strict power and area budgets (from tens to a few hundred mW and tens of mm²) [1]. The rapid evolution of AI models toward attention-based Deep Neural Networks (DNNs) (Fig. 1a) demands flexibility, motivating system-level co-design of tightly coupled clusters of programmable processors and specialized acceleration engines sharing L1 memory [2]. At the same time, the increasing heterogeneity and scale of TinyML workloads [3] are driving a shift toward multi-cluster architectures, which in turn places significant pressure on the L2 memory, shared among all clusters. Sustaining multiple clusters therefore requires high aggregate L2 bandwidth (Fig. 1b), minimizing inter-cluster contention [4], [5]. Moreover, in AI-MCU, a host core orchestrates synchronization and message passing across clusters, generating latency-critical traffic that requires fast and predictable service. This makes both average and worst-case access latency critical, motivating Quality of Service (QoS) support in the L2 memory subsystem [6].

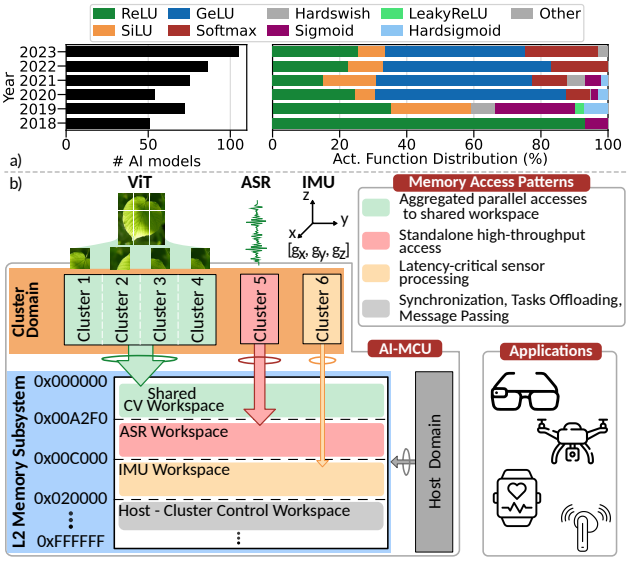


Fig. 1: (a) Growth and diversification of AI models [7] and activation functions [8] over time. b) Multi-cluster workload execution patterns and their impact on shared L2 memory.

We present Chimera[1], a flexible AI-MCU that addresses these challenges through three key innovations: **(A)** an energy-efficient Transformer Acceleration Cluster (TAC) integrating a transformer accelerator tightly coupled with fully programmable RV32IMA cores, enabling flexibility and adaptability to rapidly evolving models (Fig. 1a), achieving up to 3.1 TOPS/W; **(B)** a shared high-bandwidth AXI4-based L2 memory subsystem capable of delivering up to 563 Gb/s while mitigating inter-cluster contention, thereby enabling efficient multi-cluster workload parallelization; **(C)** a QoS-aware L2 memory architecture enabling isolation of latency-critical accesses from concurrent high-throughput traffic, achieving a 34-cycle worst-case access latency and up to 16× latency reduction compared to conventional designs. By addressing memory bandwidth scalability and contention control, Chimera enables predictable, high-performance execution of heterogeneous TinyML workloads on a low-power AI-MCU.

[1] https://github.com/pulp-platform/chimera/releases/tag/CONVOLVE-TO

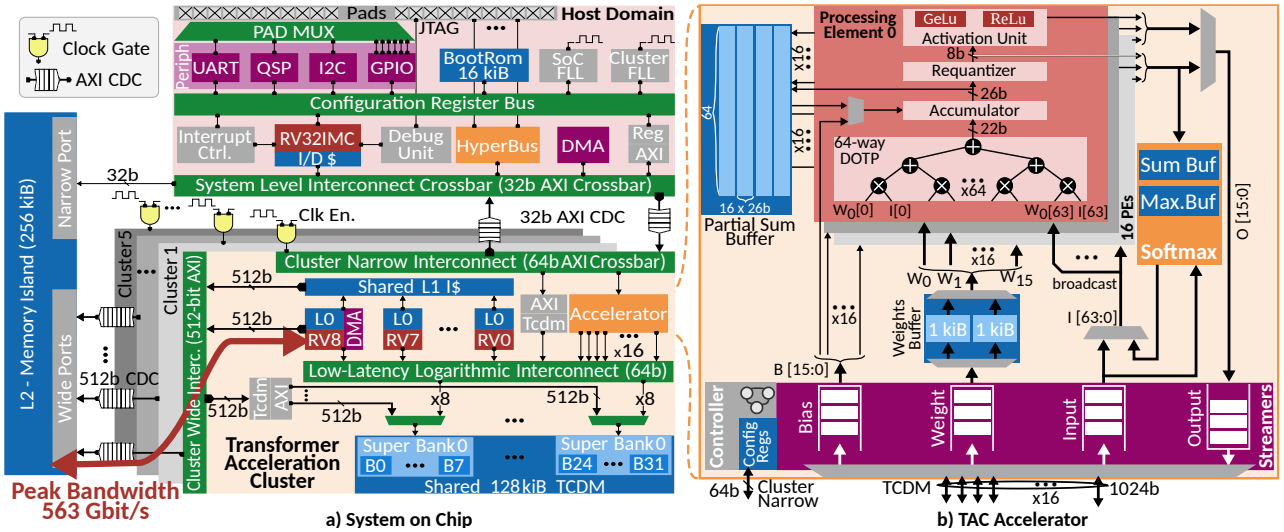
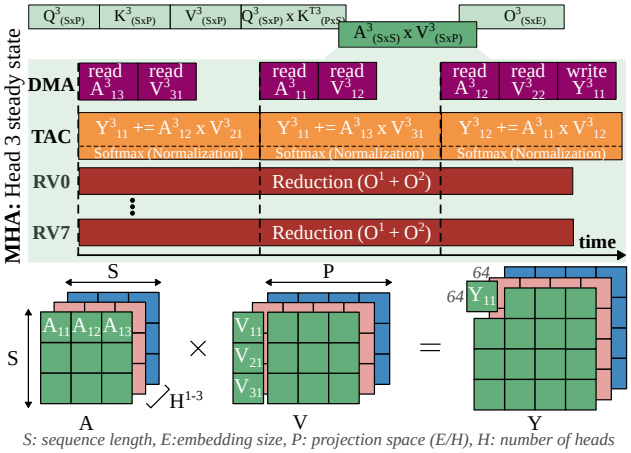


Fig. 2: (a) Architectural overview of the Chimera system-on-chip (SoC). The clusters operate in a dedicated clock domain, while the host and memory island share a common clock. AXI clock-domain crossing modules ensure synchronization, and clock-gating cells at the cluster boundary enable software-controlled clock gating. (b) TAC architecture: configuration registers are programmed via the narrow AXI interface, while streamers handle TCDM data transfers. Weights (W) are stored in a 2 KiB double-buffered memory, enabling overlap of computation and data movement. Input activations (I) are broadcast to all PEs, which compute 64-way dot products, producing 16 output elements per cycle (O). Softmax is computed on-the-fly for attention, while ReLU and GeLU are handled by the activation unit.



S: sequence length, E:embedding size, P: projection space (E/H), H: number of heads

Fig. 3: Scheduling of Multi-Head Attention (MHA) on the TAC cluster. While the accelerator computes a tile, the DMA prepares data for the next tile, and the General-Purpose (GP) cores reduce previously computed heads. The DMA can sustain computational throughput thanks to the memory island subsystem.

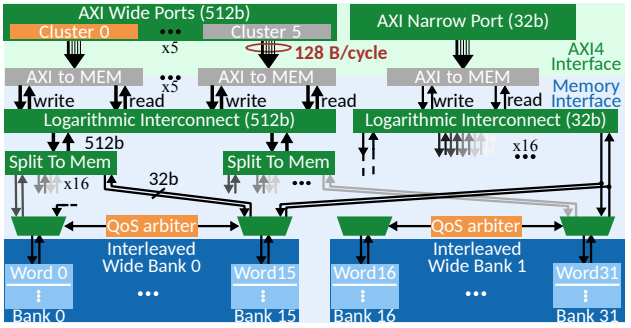


Fig. 4: L2 memory island architecture: forward arrows represent initiators, while backward arrows represent responses from target endpoints. The design features two interleaved wide banks delivering up to 128 B/cycle, along with a QoS-aware arbitration policy for latency-critical accesses.

II. Architecture

Chimera is a multi-cluster SoC (Fig. 2a) integrating seven domains to address the heterogeneous demands of TinyML signal processing. It includes a host domain, five heterogeneous clusters, and a shared L2 memory island. The host comprises an RV32IMC core responsible for system management and coordination, along with a rich peripheral subsystem including UART, I^{2}C, a HyperBus controller, and a Direct Memory Access (DMA) engine supporting data transfers between L3 memory and on-chip memories.

In this work, we focus on the Transformer Acceleration Cluster (TAC). It includes eight RV32IMA cores sharing a 128 KiB Tightly-Coupled Data Memory (TCDM), along with a 4 KiB L1 I-cache. A ninth core is dedicated to DMA management, orchestrating high-throughput transfers between the cluster and L2 memory via 512-bit ports, enabling efficient AXI4 burst transactions.

Tightly coupled with the cores, the accelerator (Fig. 2b) supports GEMM and attention mechanism using 8-bit integer quantization with minimal accuracy loss [9]. The accelerator comprises 16 Processing Elements (PEs), each operating on 8-bit weights (W) and activations (I), and computing a 64-way dot product per cycle, resulting in a peak throughput of 2048 op/cycle. Data is supplied through three streamers for inputs (I), weights (W), and biases (B), while a fourth streamer handles output write-back (O), each providing up to 128 B/cycle. To sustain this peak fetch bandwidth, the accelerator connects to the TCDM interconnect via 16 64-bit master ports. The accelerator also integrates an activation unit (8.8 kGE) in each PE, as well as a softmax engine (44 kGE) with a peak throughput of 64 softmax/cycle, operating concurrently with the PEs during attention (Fig. 3).

To support efficient data sharing and sustain high aggre-

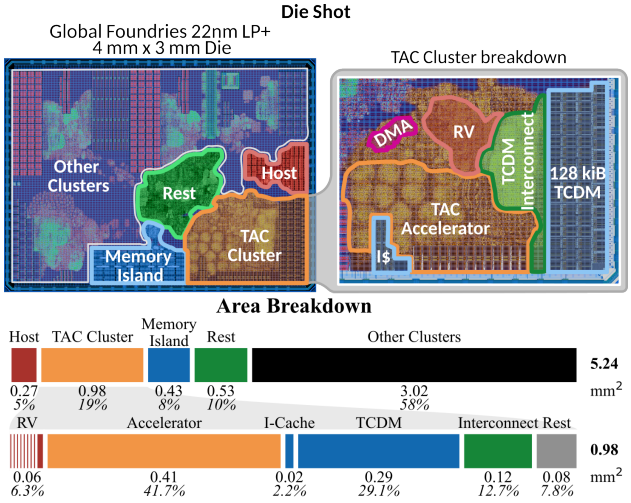


Fig. 5: Annotated chip micrograph and area breakdown. The overall die area is 12 mm^2. The silicon area evaluated in this work is 3.19 mm^2 at 60% logic area utilization.

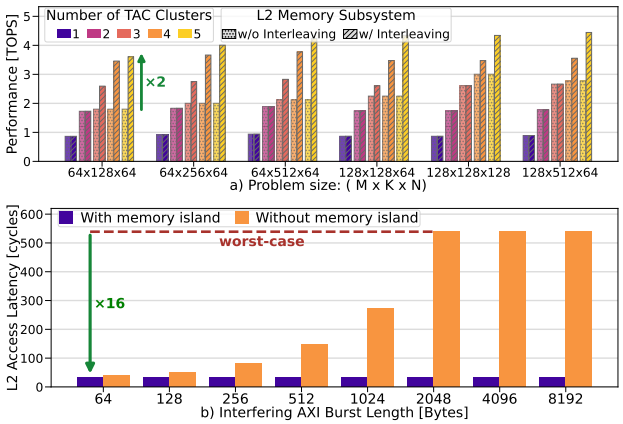


Fig. 6: (a) Simulated performance for different MATMUL sizes across multi-cluster configurations, evaluated with and without the L2 interleaved scheme. (b) Measured average L2 narrow (32-bit) access latency, under concurrent high-throughput data transfers with varying AXI4 burst lengths.

gate bandwidth across multiple clusters, Chimera features a shared 256 KiB L2 memory island (Fig. 4) with heterogeneous interfaces: 512-bit AXI4 wide interfaces for high-throughput traffic and a 32-bit AXI4 narrow interface for latency-critical messages. The wide interfaces deliver a total read/write bandwidth of 128 B/cycle per port. To sustain this bandwidth under parallel accesses, the L2 is organized into two interleaved wide banks, each 128 KiB, mitigating access conflicts and approaching the peak physical bandwidth.

However, under sustained high-throughput traffic, latency-critical messages may experience degraded QoS. To address this, the L2 supports arbitration policies including fixed priority for narrow accesses, which is effective when narrow traffic is regulated at system level, and a bounded-priority scheme to prevent starvation of wide accesses under continuous contention. This ensures low-latency service (34-cycle worst-case) for inter-cluster and host-to-cluster control traffic while maintaining high throughput for data-intensive workloads.

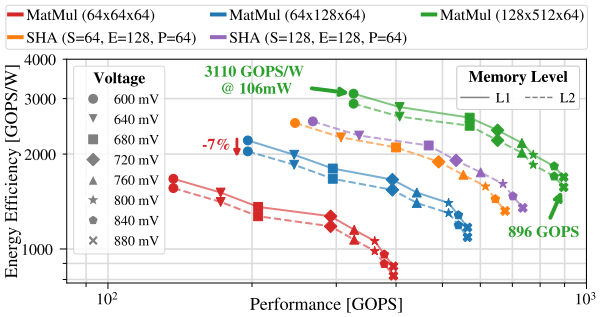


Fig. 7: Measured energy efficiency and performance for workloads executed from L1 and L2 using double buffering. Inactive clusters are clock-gated.

III. Results

Fabricated in GlobalFoundries' (GF) 22 nm LP+ technology, Chimera is designed for energy-efficient transformer-based workloads. Fig. 5 shows the annotated die micrograph. The chip occupies 12 mm^2, of which the subsystem presented in this work accounts for 3.19 mm^2 at 60% logic utilization.

To evaluate the impact of the memory island's aggregated bandwidth and interleaving scheme, matrix multiplication kernels were simulated by scaling the number of TAC clusters. As shown in Fig. 6a, beyond two active clusters, a baseline SoC without the proposed L2 subsystem is bottlenecked by inter-cluster access conflicts in the shared memory. In contrast, the proposed interleaving scheme mitigates these conflicts, enabling higher effective bandwidth than the baseline despite identical physical bandwidth, sustaining the increased throughput demand, and achieving up to 2× higher performance.

To assess the ability of the memory island to provide predictable service for latency-critical accesses under contention, QoS is evaluated by issuing 20,000 32-bit L2-to-L1 reads from the RV32IMC host core through the narrow interface, while the cluster DMA concurrently generates AXI burst reads targeting the same memory region. As shown in Fig. 6b, the baseline L2 architecture exhibits significant and burst-length-dependent latency inflation, failing to provide predictable latency. In contrast, Chimera maintains bounded and predictable host access latency under intensive high-throughput traffic from the TAC cluster, achieving up to a 16× latency reduction, confirming the effectiveness of the proposed arbitration policy.

Fig. 7 summarizes performance and energy efficiency based on the testing setup shown in Fig. 8. When executing matrix multiplication and single-head attention from L1, Chimera achieves a peak efficiency of 3.1 TOPS/W (200 MHz, 0.6 V). When the same workloads are executed from L2, the efficiency degrades by only 7%, demonstrating the effectiveness of the proposed memory subsystem. In the high-performance corner (550 MHz, 0.88 V), Chimera reaches a peak performance of 896 GOPS with a power consumption of 600 mW, within the thermal power budget of passively cooled edge devices such as wearables or palm-sized robots.

In Table I, we compare Chimera with both SoA transformer accelerators and AI-MCUs in comparable technology nodes. Compared to full MCUs architectures, Chimera

TABLE I: Comparison with State-of-the-Art (SoA).

	Ayaka JSSCC 24 [10]	EVA VLSI 25 [11]	VLSI 25 [12]	TinyVers VLSI 22 [13]	ESSERC 24 [14]	**Chimera** (This Work)
	AI Accelerator			AI-accelerated Microcontroller Unit		
Technology	28 nm	16 nm	22 nm	22 nm	18 nm	22 nm
Die Area	10.76 mm^2	16 mm^2	5.8 mm^2	6.25 mm^2	5.52 mm^2	**3.19 mm^2**
Frequency	430 MHz	1500 MHz	400 MHz	150 MHz	500 MHz	550 MHz
Voltage	0.68–1.0 V	0.49–0.9 V	0.55–0.9 V	0.4–0.9 V	0.5–0.87 V	0.6–0.88 V
Application	Transformer	Edge AI	Transformer	IoT, DNN, ML, NSA	Edge AI, IoT	**Edge AI GP, Transformer, ML**
Architecture	Transformer Accelerator	1× RISC-V + PE Array	Transformer Accelerator	1× RI5CY + ML Accel.	1× RISC-V + 128 PEs TPU	**1× RV32IMC + 9× RV32IMA + 1× TAC Accel.**
GP-Host / Prog. Accel.	✗ / ✗	✗ / ✗	✗ / ✗	✓ / ✗	✓ / ✗	✓ / ✓
L2 Aggr. Bandwidth	–	–	–	–	–	**563 Gb/s**
Precision	INT16/8	FP16	INT4/8	INT2/4/8	INT8	INT8
Peak Performance	170–**6530*** GOPS (INT8, @430 MHz)	1544 GOPS (FP16, @1.5 GHz)	920 GOPS (INT8, @400 MHz)	17.6 GOPS (INT8, @150 MHz)	–	896 GOPS (INT8, @550 MHz)
Peak Efficiency	2.22–**49.7*** TOPS/W (@0.68 V, INT8)	0.71 TOPS/W (@0.49 V, FP16)	4.46–12.52‡ TOPS/W (@0.55 V, INT8)	2.47 TOPS/W (@0.4 V, INT8)	3.25 TOPS/W† (@0.5 V, INT8)	**3.1 TOPS/W** (@0.6 V, INT8)
Area Efficiency	15.8–**607*** GOPS/mm^2	96.5 GOPS/mm^2	158.62 GOPS/mm^2	2.82 GOPS/mm^2	–	**281 GOPS/mm^2**

*The highest values are measured assuming 90% output sparsity. ‡Value obtained with one dense and one 87.5% sparse input.
† Scaling from 0.5 to 0.6 V, the efficiency is 27% lower than Chimera.

TABLE II: Full network evaluation derived from silicon measurements.

	MobileBERT	Whisper-Tiny Encoder	DINOv2-S
Model Complexity [GOP]	7.4	9.7	11.7
Throughput [1/s]	7.7–21	2.0–5.4	1.2–3.3
Energy [mJ]	9.2–16	36–72	60–118

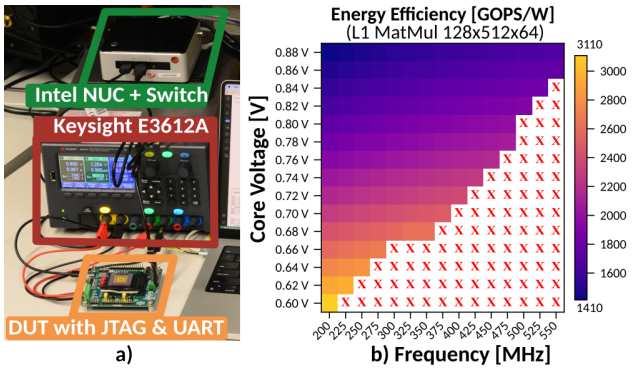


Fig. 8: a) Measurement setup: the Device Under Test (DUT) si controlled via JTAG and UART, while a programmable power supply provides and measures the SoC supply voltage and current. b) Shmoo plot for a 128 × 512 × 64 MATMUL.

achieves 1.37× higher energy efficiency and 100× higher area efficiency. Compared to pure accelerators, our architecture still achieves competitive energy efficiency for dense workloads, while providing significantly greater flexibility thanks to the tightly integrated RV processors and L2 memory hierarchy support. In addition, it achieves up to 1.8× higher area efficiency. Unlike prior AI-MCUs that primarily target CNN workloads [13], [14], this work evaluates more advanced transformer workloads in Table II, with energy per inference of 9.2/36/60 mJ for MobileBERT, Whisper-Tiny Encoder, and DINOv2-S, achieving up to 737 GOPS and 2.54 TOPS/W.

Acknowledgment

This work is funded in part by the Convolve project evaluated by the EU Horizon Europe research and innovation programme under grant agreement No. 101070374 and has been supported by the Swiss State Secretariat for Education Research and Innovation under contract number 22.00150.

References

[1] C. Silvano *et al.*, "A Survey on Deep Learning Hardware Accelerators for Heterogeneous HPC Platforms," *ACM Comput. Surv.*, vol. 57, no. 11, Jun. 2025.

[2] M. Verhelst, L. Benini, and N. Verma, "How to Keep Pushing ML Accelerator Performance? Know Your Rooflines!" *IEEE Journal of Solid-State Circuits*, vol. 60, no. 6, pp. 1888–1905, 2025.

[3] R. Aminabadi *et al.*, "DeepSpeed- Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale," in *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2022, pp. 1–15.

[4] A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer, "AI and Memory Wall," *IEEE Micro*, vol. 44, no. 3, pp. 33–39, May 2024.

[5] I. Dagli and M. E. Belviranli, "Shared Memory-contention-aware Concurrent DNN Execution for Diversely Heterogeneous System-on-Chips," in *Proceedings of the 29th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 243–256.

[6] M. Bechtel and H. Yun, "Analysis and Mitigation of Shared Resource Contention on Heterogeneous Multicore: An Industrial Case Study," *IEEE Trans. Comput.*, vol. 73, no. 7, p. 1753–1766, Jul. 2024.

[7] N. Maslej *et al.*, "Artificial intelligence index report 2025," 2025.

[8] R. Andri, E. Reggiani, and L. Cavigelli, "Flex-SFU: Activation Function Acceleration With Nonuniform Piecewise Approximation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 44, no. 11, pp. 4236–4248, 2025.

[9] G. Islamoglu *et al.*, "ITA: An Energy-Efficient Attention and Softmax Accelerator for Quantized Transformers," in *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2023, pp. 1–6.

[10] Y. Qin *et al.*, "Ayaka: A Versatile Transformer Accelerator With Low-Rank Estimation and Heterogeneous Dataflow," *IEEE Journal of Solid-State Circuits*, vol. 59, no. 10, pp. 3342–3356, 2024.

[11] J. Zhu *et al.*, "EVA: A 16mm2 1.54TFLOPS Tiled-Based Accelerator for Evolvable Edge Computing," in *2025 Symposium on VLSI Technology and Circuits*, 2025, pp. 1–3.

[12] Z. Fan, *et al.*, "A 22nm 25.08TOPS/W Multi-Task Transformer Accelerator with Mixed Precision Structured Sparsity and Two-Stage Task-Adaptive Power Management," in *2025 Symposium on VLSI Technology and Circuits*, 2025, pp. 1–3.

[13] V. Jain *et al.*, "TinyVers: A 0.8-17 TOPS/W, 1.7 µW-20 mW, Tiny Versatile System-on-chip with State-Retentive eMRAM for Machine Learning Inference at the Extreme Edge," in *2022 IEEE Symposium on VLSI Technology and Circuits*, 2022, pp. 20–21.

[14] S. Clerc *et al.*, "A 18 nm FD-SOI CMOS 6.38 mW 15 fps 8 -bit features 14.8 µJ/inference QVGA road-traffic monitoring Edge AI SoC demonstrator," in *2024 IEEE European Solid-State Electronics Research Conference (ESSERC)*, 2024, pp. 249–252.