

Toward Attention-based TinyML: A Heterogeneous Accelerated Architecture and Automated Deployment Flow

Original

Toward Attention-based TinyML: A Heterogeneous Accelerated Architecture and Automated Deployment Flow / Wiese, Philip; Islamolu, Gamze; Scherer, Moritz; Macan, Luka; Jung, Victor J. B.; Burrello, Alessio; Conti, Francesco; Benini, Luca. - In: IEEE DESIGN & TEST. - ISSN 2168-2356. - (2025), pp. 1-7. [10.1109/mdat.2025.3527371]

Availability:

This version is available at: 11583/2996569 since: 2025-01-14T09:48:11Z

Publisher:

IEEE

Published

DOI:10.1109/mdat.2025.3527371

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository









Publisher copyright

IEEE postprint/Author's Accepted Manuscript

©2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

Toward Attention-based TinyML: A Heterogeneous Accelerated Architecture and Automated Deployment Flow

Philip Wiese , *Graduate Student Member, IEEE*, Gamze İslamoğlu , *Graduate Student Member, IEEE*, Moritz Scherer , *Graduate Student Member, IEEE*, Luka Macan , *Graduate Student Member, IEEE*, Victor J.B. Jung , *Graduate Student Member, IEEE*, Alessio Burrello , *Member, IEEE*, Francesco Conti , *Member, IEEE*, Luca Benini , *Fellow, IEEE*

Abstract—One of the challenges for Tiny Machine Learning (tinyML) is keeping up with the evolution of Machine Learning models from Convolutional Neural Networks to Transformers. We address this by leveraging a heterogeneous architectural template coupling RISC-V processors with hardwired accelerators supported by an automated deployment flow. We demonstrate Attention-based models in a tinyML power envelope with an octa-core cluster coupled with an accelerator for quantized Attention. Our deployment flow enables end-to-end 8-bit Transformer inference, achieving leading-edge energy efficiency and throughput of 2960 GOp/J and 154 GOp/s (0.65 V, 22 nm FD-SOI technology).

Index Terms—Neural Networks, TinyML, Deployment, Transformers, Accelerators

I. INTRODUCTION

In recent years, Tiny Machine Learning (tinyML) has attracted much attention, bringing compute-intensive Artificial Intelligence (AI) models towards deployment on Microcontroller (MCU) class devices with power envelopes of a few milliwatts. Embedding Deep Neural Networks (DNNs) in small, low-power devices is highly relevant for numerous applications ranging from multi-modal sensing and keyword spotting to anomaly detection and smart wake-up [1]. Compared with cloud-only inference, tinyML offers lower network utilization, higher privacy, and more predictable latency. However, extreme-edge devices typically run with a tightly constrained memory budget, without fully-fledged operating systems and advanced hardware features such as Memory-Management Units (MMUs) and fully automated cache hierarchies.

One of the key research challenges is whether it is possible to build systems that respect the tight hardware and software cost and power constraints of tinyML systems while supporting the rapid advancement of models. A key consideration for addressing this research question is the trade-off between specialization and generality on the computer architecture level. Although numerous model-specific accelerators have been proposed in recent years [1], designing System-on-Chips (SoCs) that can integrate these accelerators while remaining adaptable to evolving AI models remains an open challenge, particularly under tight memory constraints. Moreover, automatically and efficiently deploying rapidly evolving DNN models, especially the increasingly popular Attention-based networks, on accelerator-enhanced MCUs remains a significant

challenge. Additionally, fast product cycles make it difficult to accommodate the time and cost associated with hand-tuning each model for deployment.

In this paper, we address what we believe to be a fundamental question for the future of tinyML: How can we move from classical perceptive AI and Convolutional Neural Network (CNN) models toward leading Attention-based Transformer models? Unlike in CNNs, complex dataflow operations like Softmax in Transformers can lead to high latency despite their low arithmetic complexity. While General Matrix Multiplication (GEMM) accelerators handle most computations in Transformer networks efficiently, the remaining operations can become a bottleneck. To address this challenge, we leverage a flexible MCU-class architectural template for efficiently integrating specialized hardware accelerators with multi-core clusters over a low-latency Tightly-Coupled Data Memory (TCDM) interconnect. At its core, we use a RISC-V (RV32) compute cluster based on the latency-tolerant Snitch core [2]. To the best of our knowledge, this is the first heterogeneous Snitch-based cluster integrating Hardware Processing Engines (HWPEs)¹, advancing beyond previous configurations which focused on instruction extension units tightly coupled to the pipeline. We introduce an extensible deployment flow based on a bottom-up DNN compiler, Deploy, that enables fast and automated End-to-End (E2E) deployment. Using this template, we integrate an extended version of the Integer Transformer Accelerator (ITA) [3] and prove our hardware-software co-design flow on Attention-based models. As a concrete use case, we showcase the E2E deployment of MobileBERT [4], DINOv2 [5], and Whisper’s encoder [6] within a power envelope of 52.0 mW (GlobalFoundries 22 nm fully-depleted silicon-on-insulator technology at 0.65 V).

The contributions of this paper are as follows:

- We propose a novel, flexible hardware-software architecture template designed to meet the dataflow and compute requirements of emerging Attention-based AI workloads. Our hardware architecture allows the co-integration of a multi-core latency-tolerant Snitch compute cluster with complex hardware accelerators over a high-bandwidth, low-latency TCDM interconnect. At the same time, our co-optimized software template facilitates efficient E2E work-

¹<https://hwpe-doc.readthedocs.io/en/latest/index.html>

load mapping. We demonstrate that our hardware-software template enables starvation-free contention for resources in the shared memory with its tunable interconnect bandwidth and the Direct Memory Access (DMA) engine. As a result, we achieve accelerator utilization of up to 85.1%.

- As a concrete use-case, we integrate ITA, a Transformer accelerator tuned for the specific dataflow of the Attention calculation, into our hardware-software template and extend Deeploy² with an accelerator model to enable automated mapping, scheduling, tiling, and code generation. We evaluate the performance through post-layout power analysis, achieving a peak performance of 741 GOP/s and energy efficiency of up to 6.35 TOP/J. The integration incurs only a 4.7 p.p. decrease in utilization compared to the standalone accelerator, demonstrating the low overhead of our template.
- We showcase the capability of our hardware-software template to support a range of Attention-based tinyML models, including MobileBERT, DINOv2, and Whisper’s encoder. Our flow unlocks the potential for collaborative execution between the cluster and the hardware accelerator, which optimizes performance and energy efficiency and prevents resource starvation. By enabling this collaborative execution, we significantly enhance E2E inference energy efficiency by $102\times$ compared to inference without the accelerator, achieving an E2E throughput of up to 154 GOP/s and energy efficiency of 2.96 TOP/J.

II. THE CHALLENGES OF TINYML ACCELERATION

A. HW Integration Challenges for Attention-based Networks

Over the years, several approaches to integrating hardware accelerators into SoCs were proposed, varying in their degree of coupling to the SoC’s processor.

A well-developed approach relies on closely coupling the accelerator with the processor cores through instruction-set extensions [7]. While this approach enables ample flexibility in workload mapping, it is inadequate for Attention accelerators that require large bandwidth. In fact, instruction extensions are limited by the core’s load/store interface, the bandwidth and size of the register file, and instruction fetch bandwidth.

On the other end of the spectrum is the loosely coupled integration of accelerators with internal private memory [8]. While this approach eliminates memory access contention during inference, it requires a large in-accelerator and fully private integrated memory to store the intermediate tensors generated for Attention. This causes large area requirements, which increase the cost of the accelerator. It also hinders collaboration between different engines, as data must be moved explicitly between memory hierarchy levels with a significant energy overhead. An interesting middle ground between these two extremes is to couple the accelerator and cores through shared memory [8]. Unlike private memory solutions, this approach facilitates data exchange between the accelerators and cores. This is a key feature for Attention-based networks since it allows cores to perform auxiliary operations easily without memory

copy overheads. These operations vary significantly across different model variants, often preventing hardware acceleration.

In this work, we propose a novel architectural template, integrating a cluster of RISC-V cores with an accelerator over shared L1 memory. We show that our proposed design enables close interaction between the cluster cores and the accelerator, supporting emerging and evolving variations of non-linearities and normalization layers found in Attention-based models while exploiting the accelerator for supported operators.

B. TinyML Software Deployment Challenges

Deploying Transformers at the extreme edge on devices with hardware accelerators comes with many difficulties as they require significant software effort to unlock the performance and efficiency of the accelerators. First and foremost, tinyML devices have highly constrained on-chip memory, in the order of MiB, and no operating systems. Hence, one must *tile* layers to process tensors from the lowest level of the memory hierarchy. Moreover, these systems often feature software-managed scratchpad memory hierarchies. Thus, explicit and uncached DMA transfers are required to transfer tiled tensors. Furthermore, static memory allocation is crucial to guarantee conflict-free memory transfers.

While several code generation tools for CNNs have been demonstrated [1], most do not generalize to Attention-based models. While CNNs use few branches in their dataflow graphs and therefore do not require sophisticated memory allocation strategies, the highly parallel and branching structure of Attention-based networks requires novel lifetime analysis and tiling strategies to effectively tile and schedule their execution.

III. ARCHITECTURE TEMPLATE

In this Section, we describe a flexible architecture template, shown in Figure 1, that combines multiple Digital Signal Processing (DSP) optimized RISC-V cores into a compute cluster and facilitates the integration of newly developed hardware accelerators using the HWPE infrastructure and automated deployment. Compared to a single-core system, this enables efficient operation through higher performance and parallelism and enhances adaptability and scalability. The HWPE interface developed for the Parallel Ultra Low Power (PULP) platform facilitates the integration of accelerators with multi-core compute cluster into a shared memory cluster. Our template integrates the area-efficient Snitch cores, occupying 22 kGE each [2]. Snitch is a single-stage, in-order core implementing integer base RV32I, RV32M subset for integer multiply/divide instructions, and standard atomic instruction extension RV32A. Unlike CV32E40P cores used in other PULP-derived clusters³, Snitch cores are significantly smaller (-56%) and provide a decoupled memory interface, allowing latency-tolerant memory access by pipelining multiple loads and stores.

We couple the cores and accelerators through the shared interleaved L1 TCDM to facilitate energy-efficient data exchange between the compute elements. This is especially crucial for rapidly evolving Attention-based networks as various auxiliary operations need to be computed on the cluster

²<https://github.com/pulp-platform/Deeploy>

³<https://docs.openhwgroup.org/projects/cv32e40p-user-manual>

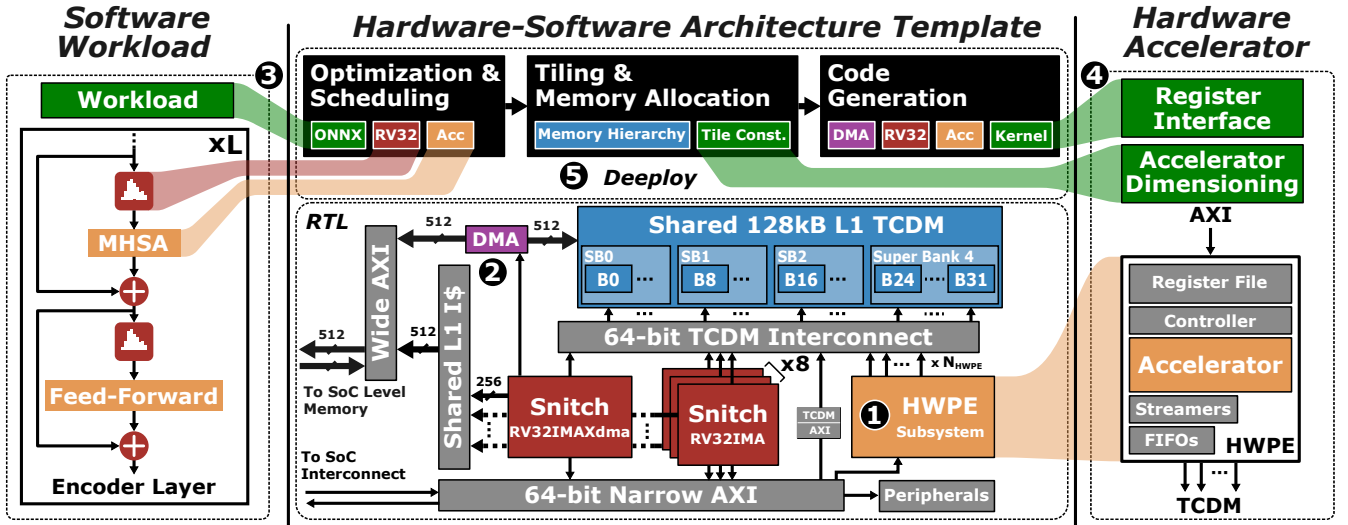


Fig. 1. Overview of the Hardware-Software Architecture Template. The flexible template allows modular integration of accelerators into an SoC and deployment of different workloads with Deeploy. The workflow is as follows: ① Integrate an accelerator as an HWPE engine, a configurable interface designed for efficient integration of memory-coupled accelerators, enabling streamlined data transfer and control between the accelerator and shared memory. ② Ensure sufficient bandwidth for the accelerator by tuning the wide AXI interconnect, allowing high-bandwidth access to L2 memory via the DMA. ③ Configure the operator mapping in Deeploy and provide the workload as an ONNX graph. ④ Define the tiling constraints according to the accelerator buffer and datapath sizes and provide minimal kernel templates to control the accelerator via a register interface. ⑤ Use Deeploy to perform automated graph optimization and scheduling, to co-optimize operator tiling and static memory allocation, and to generate C code. This code orchestrates memory transfers using the DMA and coordinates execution on the compute cores and the accelerator.

while the majority of the computation is conducted on the accelerator. To reduce banking conflicts and provide the high bandwidth Attention accelerators need, we use 32 banks with 4 KiB each, resulting in a total capacity of 128 KiB. The multi-banked memory makes it unnecessary to attach additional private memory to the accelerator, as data can be accessed by both the accelerator and the cluster’s cores simultaneously. We use a 64-bit TCDM interconnect, which is implemented as a combinatorial crossbar, resulting in single-cycle latency in the absence of conflicts with 256 B/cycle bandwidth towards the L1. Each core has one master port with decoupled request and response path connected to the TCDM interconnect, and the HWPE subsystem features a parametric number N_{HWPE} of master ports to allow the integration accelerators.

The cluster includes two parametrizable AXI interconnects: a wide crossbar with a $D_{AXI,W}$ bit data width and a narrow crossbar with a $D_{AXI,N}$ bit data width. The wide AXI interconnect is used to load instructions into the shared 8 KiB instruction cache and to transfer data from and to the SoC level memory system in conjunction with the DMA. The narrow AXI interconnect is intended to connect to the SoC interconnect to attach peripherals and communicate with a host system. Moreover, one Snitch core is coupled with a DMA to manage data movements within the cluster, facilitating double buffering to maintain high accelerator utilization.

A. HWPE Subsystem

The HWPE template provides three modules: a *controller*, one or multiple *streamers*, and the *engine*. The *controller* is the interface between the cores in the cluster and the accelerator. It has a Finite State Machine (FSM) specific to the engine to govern the operation of the accelerator and a memory-mapped

register file to keep parameters for the accelerator. The register file can hold a sequence of multiple *tasks* that can be programmed by any core in the cluster through the controller interface over the narrow AXI interconnect. A *task* represents a set of configuration values used by the accelerator. The *streamers* act as a special-purpose low-cost DMA to load and store data from the shared TCDM. Finally, the *engine* contains a hardware accelerator that accepts the streamer’s data and the controller’s configuration.

HWPE allows connecting accelerators seamlessly to PULP clusters and makes programming straightforward over the peripheral interface accessible via AXI. Three steps are necessary to integrate an accelerator into the HWPE subsystem. First, the required number of streamers must be instantiated in accordance with the accelerator’s data ports. Next, the streamers must be connected with the accelerator’s data ports and the TCDM interconnect. Finally, an FSM controlling the accelerator and streamers must be implemented.

HWPE provides two types of streamers: one for input, *source streamers* and one for output, *sink streamers*. The streamers utilize a simple valid-ready handshake protocol on the accelerator side, ensuring compatibility with most accelerators. Additionally, HWPE includes first-in, first-out buffers (FIFOs) on both the TCDM and accelerator sides, which can be instantiated and sized according to the specific needs of the accelerator and cluster. We time-multiplex multiple streamers to a multi-port interface with N_{HWPE} ports and connect to the TCDM interconnect.

The final step of integrating an accelerator into the HWPE involves designing an FSM to control both the accelerator and the streamers. We use a *controller* that supports a programmable multi-context register file, allowing the cores to offload the next *task* while the accelerator runs, thereby

hiding configuration latency. The FSM designed around the control slave is straightforward: it reads the configuration for the accelerator from the register file, transfers it to the engine, and configures the streamers accordingly.

B. Neural Network Deployment Framework

To execute Transformer models on the proposed architectural template, we integrate our hardware template in the Deeploy compiler [9], which maps neural networks to user-defined, platform-specific C code kernel templates. Deeploy is a DNN compiler that offers architecture-agnostic tinyML optimizations like double-buffering, memory-aware operator tiling, DMA-aware code generation, and fully static offline memory layout generation. These features allow us to accommodate the custom tiling required for operators exclusively present in Transformer networks.

In this way, Deeploy generates code to offload supported DNN operators onto accelerators while providing highly optimized fallback kernel implementations for unsupported operators on the cluster. This bottom-up approach guarantees that emerging DNNs operators can be mapped to our general-purpose cores while fully leveraging integrated accelerators for their supported operators. This is especially useful when considering the numerous variants of Attention-based models, which contain the same Attention mechanism but have slightly different activation or normalization functions.

To integrate a new HWPE accelerator, Deeploy only requires a minimal accelerator model; first, the accelerator model must specify the geometrical tiling constraints for operators it can run. Second, the model must provide minimal arithmetic templates for running each supported operator. All other necessary performance optimizations, including memory-aware operator tiling, static memory layout generation, double-buffering code generation, and DMA-aware memory transfers, are inserted by Deeploy automatically.

By integrating a model of the hardware template with Deeploy, we propose a low-overhead, adaptable hardware-software architecture template that minimizes the development effort for both hardware and software integration while meeting the strict requirements of extreme edge Attention-based model deployment.

IV. IMPLEMENTATION

As a concrete implementation of our template, we show a platform that couples a cluster with 8+1 RV32IMA Snitch cores with the Integer Transformer Accelerator (ITA) [3]. The ITA accelerator enables the acceleration of 8-bit GEMM and the more complex multi-head Attention (MHA) present in Transformer networks. ITA used in this work is an extended version of the accelerator presented in [3], featuring additional functionality through the inclusion of a partial sum buffer and an activation unit supporting ReLU and GeLU. Furthermore, it is wrapped with HWPE components.

A. Integer Transformer Accelerator (ITA)

ITA is an accelerator for encoder-only Transformer models and performs efficient inference in 8-bit arithmetic, using an

integer-only Softmax approximation. Figure 2 shows the architecture of ITA. At the core of ITA, there are N dot product units that compute the dot product between two vectors of length M .

ITA integrates a Softmax approximation, referred to as *ITAMax*, that operates on integer values in a streaming mode. This enables computing Softmax on the fly. Softmax is defined as

$$\text{Softmax}(\mathbf{x})_i = \frac{e^{x_i - \max(\mathbf{x})}}{\sum_{j=1}^n e^{x_j - \max(\mathbf{x})}} \quad (1)$$

and normalizes the input matrix row-wise, transforming them into probabilities. This is used in Transformers to calculate the Attention $\mathbf{A} \times \mathbf{V}$ with

$$\mathbf{A} \times \mathbf{V} = \text{Softmax}(\mathbf{Q} \times \mathbf{K}^T) \times \mathbf{V} \quad (2)$$

The *ITAMax* unit has three stages of operation as illustrated in Figure 2. The first Denominator Accumulation (DA) stage operates on the 8-bit quantized dot product results from the $\mathbf{Q} \times \mathbf{K}^T$ matrix multiplication. It determines the maximum of the partial row results and accumulates the denominator of the Softmax with the current maximum. The current maximum and the accumulated denominator are stored in buffers. At every iteration, if the local row maximum differs from the previous one, the partial sum is renormalized, and the global maximum is updated.

Once *ITAMax* processes the entire row and accumulates the denominator with the global maximum of the row, it inverts the denominator in the Denominator Inversion (DI) stage and stores it internally. The Element Normalization (EN) stage only starts when the post-Softmax activations are required as input to ITA in the next matrix multiplication ($\mathbf{A} \times \mathbf{V}$). This stage normalizes the values from the $\mathbf{Q} \times \mathbf{K}^T$ calculation on the fly to produce \mathbf{A} .

With this unique dataflow, *ITAMax* performs Softmax without additional latency and data fetching from the L1 memory with a low area and power overhead. Since ITA integrates a datapath for single-head Attention, MHA must be calculated sequentially head-by-head. Therefore, ITA operates on a single head at a time and computes the partial output projection for each head. The partial outputs of each head need to be summed by the external cluster.

Additionally, ITA integrates activation units that fully operate in integer arithmetic. The activation unit has three modes of operation: Identity, ReLU, and GeLU, which can be selected for each computation via the configuration interface of HWPE. For the integer approximation of GeLU, we use the i-GeLU [10] performed in D -bit and quantized the results to 8-bit. This allows using ITA as a GEMM accelerator with activation functions accelerated in hardware.

B. Accelerator Integration

For ITA, we use $N=16$ dot product units with a $D=26$ -bit accumulator to support matrix dimensions up to 512 and a vector length of $M=64$. We choose this configuration to exploit the memory-side bandwidth the TCDM offers. As ITA features three input ports (input, weight, bias) and one output port, three input streamers and one output streamer are required.

As the four streamers are multiplexed in time, ITA requires 128 B/cycles of maximum bandwidth to fetch two input vectors

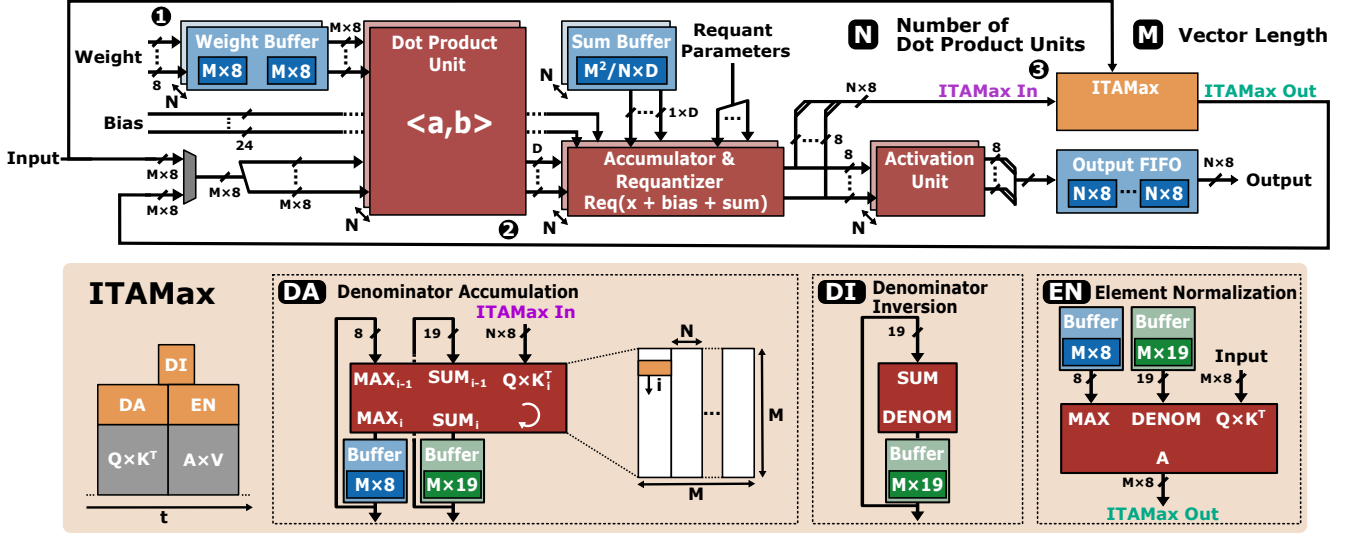


Fig. 2. Architecture of the Integer Transformer Accelerator (ITA). ITA combines an output stationary dataflow with a local weight stationary dataflow and streaming Softmax operation to achieve high data reuse and minimal memory interaction. ① Weights are stored in a double-buffered weight memory to fetch the next set of weights while performing computation with the current set of weights. ② Inputs are fetched via streamers and passed through the ITAMax module during $A \times V$ step. ③ While $Q \times K^T$ is computed, the ITAMax module operates on the outputs to accumulate the denominator. ITAMax operates in three stages: **DA** Find the local maximum and compare it with the previous maximum stored in the buffer, accumulate the denominator of the Softmax using the current maximum and normalize the previous sum if the maximum is changed. **DI** After the accumulation, the denominator is inverted and saved to the same buffer. **EN** Inputs for $A \times V$ step are normalized using the saved maximum and inverted denominator.

per cycle; therefore, we use 16 master ports on the TCDM interconnect for the HWPE subsystem. To produce one output tile, ITA takes at least 256 cycles and the DMA needs to fetch at most two 64×64 8-bit inputs/weights, 64 24-bit bias values and write back 64×64 8-bit outputs from and to the L2 memory. This results in a worst-case average bandwidth of 48.75 B/cycles towards the SoC memory. Consequently, we use a 512-bit wide data AXI interconnect to provide enough bandwidth for the instructions cache and ITA. Moreover, we use 64-bit for the narrow AXI interconnect to enable the integration of the cluster into a 64-bit host system. Finally, in ITA, we use a dual-context register file that can be programmed via the narrow 64-bit AXI interconnect. As the HWPE *Controller* uses the peripheral interface, we place an adapter between the AXI bus and the module.

C. Physical Implementation

To evaluate our architecture in a tinyML-friendly technology node, we implemented the complete Snitch cluster with an extended version of the ITA accelerator in GlobalFoundries' 22 nm FDX fully-depleted silicon-on-insulator (FD-SOI) technology, targeting an operating frequency of 500 MHz under typical conditions (TT, 0.8 V, 25 °C), and 425 MHz in the energy-efficient core voltage configuration (TT, 0.65 V, 25 °C). The extended design includes a partial sum buffer, an activation unit, and the HWPE components. The complete cluster requires 0.991 mm² (5 MGE) with the HWPE subsystem occupying 39.3% of the total area. The longest paths of the design are located between the input to the output of the dot product units in the HWPE, within the DMA, and the instruction cache to the data mover core with gate delays of 12, 11, and 11, respectively.

D. Neural Network Deployment

To extend Deeploy with our architecture template, including the cluster and ITA, the mapping process of ITA-compatible operators is implemented in a multi-step approach. Deeploy starts by matching an MHA pattern and fuses it to form a monolithic node in the graph. This node is then split along the head dimension to map the MHA operator head-by-head on ITA. Finally, a head accumulation layer is inserted at the end, which runs on the cluster cores.

As described in Section III-B, we extend Deeploy with a model for ITA to support HW-specific optimizations. To solve the tiling problem, we specify geometrical tiling constraints to ensure all inputs and outputs have shapes compatible with ITA's requirements. In the kernel, we preprogram the next tile using the dual-context register file and configure ITA to load the weights for the next step in the current one. This enables us to achieve a fully double-buffered dataflow without starvation.

To the best of our knowledge, this is the first deployment flow that supports the E2E acceleration of Attention-based Transformers at the edge.

V. RESULTS

To measure the power consumption and latency of deployed workloads on our design, we perform cycle-accurate post-layout simulation of the entire cluster using Siemens QuestaSim for latency and throughput evaluation at 425 MHz and post-layout gate-level simulations for power measurement under typical conditions (TT, 0.65 V, 25 °C). We choose the 0.65 V operating corner to maximize energy efficiency. Our simulation setup accounts for latency and energy costs of memory transfers between the L1 and the system's background memory via the DMA, programming of the accelerator and cores, and

TABLE I
END-TO-END NETWORK PERFORMANCE METRICS AND COMPARISON TO DNNs ON COMMERCIAL TINYML DEVICES

Metric	Unit	Ours		Commercial Devices			
		Multi-Core	Multi-Core + ITA	Syantiant NDP120 [‡] [11]	AlifSemi E3 [§]	GreenWaves GAP9 ^{*‡} [11], [12]	
Throughput	[GOp/s]	0.74	56-154	2-7	2-45	10-60	
Energy Efficiency	[GOp/J]	28.9	1600-2960	280-400	50-560	150-650	
Power	[mW]	26.0	35.2-52.0	-	-	-	

Metric	Unit	MobileBERT ^a		DINOv2-Small ^b		Whisper-Tiny Encoder ^c	
		Multi-Core	Multi-Core + ITA	Multi-Core	Multi-Core + ITA	Multi-Core	Multi-Core + ITA
Energy per Inference	[mJ/Inf]	164	1.60	407	7.31	340	5.55
Inference per Second	[Inf/s]	0.16	32.5	0.06	4.83	0.08	6.52

[‡] MobileNetV1(x0.25) with 28 MOp

[§] MicroNet Medium, MobileNetV2 1.0, Yolo-Fastest v4, Tiny Wav2letter Pruned, <https://alifsemi.com/>

^{*} TinyissimoYOLO

^a 4.74 GOP per inference with sequence length $S=128$

^b 11.7 GOP per inference with sequence length $S=241$

^c 9.74 GOP per inference with sequence length $S=512$

execution of the operators, both on the cluster and ITA. In the following sections, we profile representative microbenchmarks and the execution of three different Transformer networks. Finally, we compare our results with state-of-the-art MCU-class heterogeneous SoCs for tinyML.

A. Microbenchmarking Result

We analyze the performance and efficiency of GEMM and the more complex Attention operations and compare the multi-core cluster without any accelerator with the ITA integrated cluster. Our heterogeneous cluster achieves a throughput of 741 GOP/s and energy efficiency of 5.42 Top/J in GEMM computation, corresponding to 986 \times and 188 \times improvement respectively compared to the cluster without ITA with a peak accelerator utilization of 85.1%. Running single-head Attention operation offers an even higher performance improvement of more than 3 orders of magnitudes and a 901 \times better energy efficiency resulting in 663 GOP/s and 6.35 Top/J with 74.9% accelerator utilization. The standalone accelerator achieves a slightly higher utilization of 79.6%, with the integration into the template incurring only a small decrease of 4.7 p.p.. This demonstrates that the template has minimal impact on the accelerator utilization. This trend can be attributed to the efficient Softmax implementation in ITA, which does not add latency and thus avoids bottlenecking the overall efficiency.

B. End-To-End Deployment Results

To benchmark the execution of a complete model, we quantize MobileBERT⁴, DINOv2⁵ and Whisper⁶ encoder using the QuantLib⁷ library to perform 8-bit full integer inference. Due to the extensive simulation time, we measure each layer separately and sum their execution times to extrapolate to the entire network. Table I display the E2E

⁴ $S = 128, E = 128, P = 64, H = 4, N = 24, d_{ff} = 512$ (Sequence Length, Embedding Size, Projection Dimension, Attention Heads, Layers, Feed-Forward)

⁵ $S = 241, E = 384, P = 64, H = 6, N = 12, d_{ff} = 1536$

⁶ $S = 512, E = 384, P = 64, H = 6, N = 4, d_{ff} = 1536$

⁷<https://github.com/pulp-platform/quantlib>

results for two scenarios: multi-core cluster without the accelerator and multi-core cluster with the ITA accelerator.

In the scenario with a multi-core cluster, using ITA improves throughput up to 208 \times at 102 \times higher energy efficiency.

C. Comparison with the State-of-the-art

To compare our work with the state-of-the-art in tinyML computer architectures, we present the throughput and energy efficiency for various devices in Table I. Due to the lack of E2E benchmarks for Transformers on similar devices, we compare against CNNs instead. It is important to note that Transformers pose a greater challenge for accelerators due to their complex dataflow and computational demands.

The Syntiant NDP120⁸ MCU implemented in UMC 40 nm ULP technology uses the Syntiant Core 2 tensor processor coupled with an Arm Cortex M0 processor and a HiFi-3 DSP. It achieves up to 7 GOP/s at 400 GOP/J in *MLPerf Tiny Inference* on MobileNetV1 [11]. We also compare with the Ensemble E3 AI MCU from Alif Semiconductor⁹ which couples Ethos-U55 Machine Learning (ML) processors with ARM Cortex M55 processors. Depending on the network it achieves up to 45 GOP/s at 560 GOP/J. Compared to both devices, we achieve at least 3.4 \times more throughput with a 5.3 \times higher energy efficiency.

A comparison with a very similar architecture is possible against GreenWaves GAP9 MCU containing the NE16 neural engine. The SoC implemented in 22 nm technology contains a fabric controller and a compute cluster with nine RISC-V cores and 128 kB shared L1 memory. In the *MLPerf Tiny Inference* benchmark on MobileNetV1 it achieves 25 GOP/s at 480 GOP/J while Moosmann et al. [12] report better numbers with up to 60 GOP/s at 650 GOP/J for a different network. In comparison, we achieve 2.6 \times more throughput and 4.6 \times higher energy efficiency even though we deploy a more complex network.

VI. CONCLUSION

We presented a flexible hardware-software architecture template, enabling collaborative accelerated execution of

⁸<https://www.syntiant.com/hardware>

⁹<https://alifsemi.com/>

emerging Attention-based workloads that can be easily extended for the demands of future networks. By integrating our hardware template in Deeploy, we demonstrate a flexible deployment flow capable of efficiently mapping both accelerator-specific and generic DNN operators on our target architecture. We demonstrate the first E2E deployment of multiple Transformer-based encoder models on a parallel heterogeneous accelerator-enhanced MCU. Our implementation, which leverages ITA for computing the MHA and Linear layers and the cluster cores for auxiliary operators, achieves state-of-the-art throughput of 154 GOP/s with an energy efficiency of 2.96 TOP/J. This enables inference rates of 32.5 Inf/s at 1.60 mJ/Inf for MobileBERT, 4.83 Inf/s at 7.31 mJ/Inf for DINOv2-Small, and 6.52 Inf/s at 5.55 mJ/Inf for encoder block of Whisper.

ACKNOWLEDGMENT

We thank Andrei Deaconeasa, Maximilian Coco, and Timon Fercho for their valuable contributions to the research project. This work is supported in part by CONVOLVE (g.a. 101070374) and NeuroSoC (g.a. 101070634) projects under the European Union’s Horizon research and innovation programme, and TRISTAN (g.a. 101095947) project funded by Chips JU.

REFERENCES

- [1] Y. Abadade *et al.*, “A Comprehensive Survey on TinyML,” *IEEE Access*, vol. 11, pp. 96 892–96 922, Jul. 2023.
 - [2] F. Zaruba *et al.*, “Snitch: A Tiny Pseudo Dual-Issue Processor for Area and Energy Efficient Execution of Floating-Point Intensive Workloads,” *IEEE Transactions on Computers*, vol. 70, no. 11, pp. 1845–1860, Nov. 2021.
 - [3] 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)*, Aug. 2023.
 - [4] Z. Sun *et al.*, “MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky *et al.*, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 2158–2170.
 - [5] M. Oquab *et al.*, “DINOv2: Learning Robust Visual Features without Supervision,” *Transactions on Machine Learning Research*, Jul. 2023.
 - [6] A. Radford *et al.*, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 28 492–28 518, iSSN: 2640-3498.
 - [7] E. Cui *et al.*, “RISC-V Instruction Set Architecture Extensions: A Survey,” *IEEE Access*, vol. 11, pp. 24 696–24 711, Feb. 2023.
 - [8] E. G. Cota *et al.*, “An analysis of accelerator coupling in heterogeneous architectures,” in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, Jun. 2015, pp. 1–6.
 - [9] M. Scherer *et al.*, “Deeploy: Enabling Energy-Efficient Deployment of Small Language Models on Heterogeneous Microcontrollers,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 11, pp. 4009–4020, Nov. 2024.
 - [10] S. Kim *et al.*, “I-BERT: Integer-only BERT Quantization,” in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 5506–5518.
 - [11] C. Banbury *et al.*, “MLPerf Tiny Benchmark,” *arXiv*, no. arXiv:2106.07597, Aug. 2021.
 - [12] J. Moosmann *et al.*, “Ultra-Efficient On-Device Object Detection on AI-Integrated Smart Glasses with TinyissimoYOLO,” *arXiv*, no. arXiv:2311.01057, Nov. 2023.
- Philip Wiese** received the B.Sc. and M.Sc. degree in electrical engineering and information technology from ETH Zürich in 2021 and 2023, respectively, where he is currently pursuing a Ph.D. degree at the Integrated Systems Laboratory. His research interests include machine-learning compilers and digital low-power design. Contact him at wiesep@iis.ee.ethz.ch.
- Gamze İslamoğlu** received her B.Sc. degrees in Electrical and Electronics Engineering, and Physics from Boğaziçi University in 2020, and M.Sc. degree in Electrical Engineering and Information Technology from ETH Zürich in 2022. She is currently pursuing a Ph.D. degree at the Integrated Systems Laboratory at ETH Zürich. Her research interests include hardware accelerators for machine learning and heterogeneous multicore SoCs. Contact her at gislamoglu@iis.ee.ethz.ch.
- Moritz Scherer** received the B.Sc. and M.Sc. degree in electrical engineering and information technology from ETH Zürich in 2018 and 2020, respectively, where he is currently pursuing a Ph.D. degree at the Integrated Systems Laboratory. His research interests include ultra-low power and energy-efficient circuits and embedded design for machine learning. Contact him at scheremo@iis.ee.ethz.ch.
- Luka Macan** received the B.Sc. and M.Sc. degree from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia in 2017 and 2019, respectively. He is currently pursuing a Ph.D. degree at the University of Bologna. His research interests include machine learning on embedded systems and hardware accelerators. Contact him at luka.macan@unibo.it.
- Victor Jean-Baptiste Jung** received his Bachelor’s degree in Computer Science and Engineering Physics from Juniata College, and the Master’s degree in Computer Science from the Institut Supérieur de l’Electronique et du Numérique of Lille (ISEN Lille) in 2022. He is currently pursuing a Ph.D. degree at the Integrated Systems Laboratory at ETH Zürich. His current research interests include the efficient deployment of ML models on microcontrollers and quantization. Contact him at jungvi@iis.ee.ethz.ch.
- Alessio Burrello** received his M.Sc. and Ph.D. degrees in Electronic Engineering at the Politecnico di Torino, Italy, and the University of Bologna, respectively, in 2018 and 2023. He is currently working as a research assistant at Politecnico di Torino. His research interests include parallel programming models for embedded systems and hardware-oriented deep learning. Contact him at alessio.burrello@polito.it.
- Francesco Conti** received the Ph.D. degree in electronic engineering from the University of Bologna, Italy, in 2016. He is currently a Tenure-Track Assistant Professor with the DEI Department, University of Bologna. His research interests include hardware acceleration in ultra-low power SoCs for artificial intelligence applications. Contact him at f.conti@unibo.it.
- Luca Benini** holds the Chair of Digital Circuits and Systems at ETH Zürich and is a Full Professor with the Università di Bologna. He is a Fellow of the ACM and the IEEE and a member of the Accademia Europaea. His research interests include energy-efficient computing systems and machine-learning hardware. Contact him at lbenini@iis.ee.ethz.ch.