

# T-SAR: A Full-Stack Co-design for CPU-Only Ternary LLM Inference via In-Place SIMD ALU Reorganization

Hyunwoo Oh, KyungIn Nam, Rajat Bhattacharjya, Hanning Chen, Tamoghno Das, Sanggeon Yun, Suyeon Jang, Andrew Ding, Nikil Dutt, and Mohsen Imani  
 Department of Computer Science, University of California, Irvine  
 Email: {hyunwoo, m.imani}@uci.edu

**Abstract**—Recent advances in LLMs have outpaced the computational and memory capacities of edge platforms that primarily employ CPUs, thereby challenging efficient and scalable deployment. While ternary quantization enables significant resource savings, existing CPU solutions rely heavily on memory-based lookup tables (LUTs) which limit scalability, and FPGA or GPU accelerators remain impractical for edge use. This paper presents T-SAR, the first framework to achieve scalable ternary LLM inference on CPUs by repurposing the SIMD register file for dynamic, in-register LUT generation with minimal hardware modifications. T-SAR eliminates memory bottlenecks and maximizes data-level parallelism, delivering  $5.6\text{--}24.5\times$  and  $1.1\text{--}86.2\times$  improvements in GEMM latency and GEMV throughput, respectively, with only 3.2% power and 1.4% area overheads in SIMD units. T-SAR achieves up to  $2.5\text{--}4.9\times$  the energy efficiency of an NVIDIA Jetson AGX Orin, establishing a practical approach for efficient LLM inference on edge platforms.

**Index Terms**—Large Language Models, SIMD, Instruction Set Architecture, Quantization, Ternary LLM.

## I. INTRODUCTION

Large Language Models (LLMs) have become ubiquitous today, with numerous applications, including those in coding assistants, document analysis, and interactive conversational interfaces across consumer and enterprise systems [1]. However, LLMs require substantial resources for inference, with billions of parameters driving extensive matrix operations and creating significant latency during autoregressive generation, where each token requires a full model forward pass [1].

Traditionally, LLMs have been deployed on cloud servers with high-power GPUs, NPUs, or specialized accelerators to handle their extreme compute and memory demands (e.g.,  $>140$  GB memory for Llama2-70B [2]). Increasingly, however, there is a need to run LLMs directly on the edge for scenarios such as coding copilots with proprietary source code, document analysis of confidential business data, and personalized assistants handling sensitive information [3]. In these settings, reliance on cloud computing is often infeasible due to intellectual property concerns, data privacy regulations like GDPR and HIPAA [4], limited network connectivity, or prohibitive cloud costs for continuous inference workloads.

Thus, to enable standalone LLM deployment on edge devices at lower cost, several techniques have been proposed, including pruning [5]–[7], quantization [8]–[11], knowledge distillation [12]–[14], and weight binarization [15]–[17].

Within quantization, ternary quantization has emerged as a particularly promising approach [18]–[20]. By constraining weights to  $\{-1, 0, 1\}$ , it achieves  $8\times$  memory compression

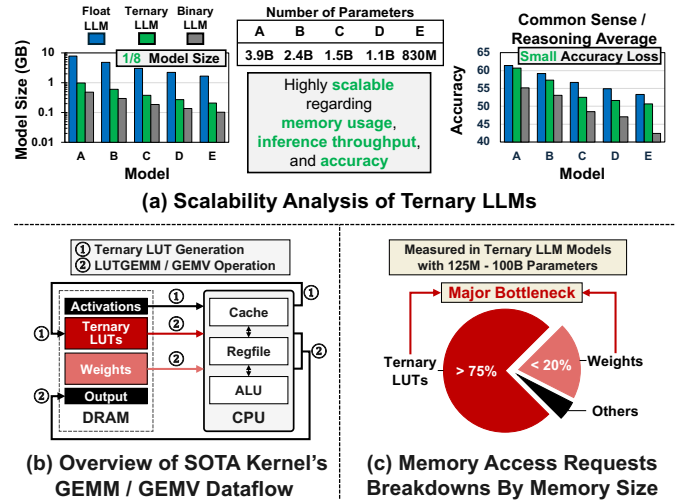


Fig. 1: Motivation for scalable ternary LLM acceleration. (a) Ternary LLMs provide  $8\times$  size reduction with minimal accuracy loss, making them suitable for edge deployment. (b) GEMM/GEMV dataflow: SOTA LUT-based kernels store TLUTs in DRAM, causing frequent memory access requests. (c) Memory access breakdown: TLUTs dominate system memory requests—over 75%—across models from 125M to 100B parameters, creating a major bottleneck for CPU inference.

(Fig. 1(a)) while maintaining 93–99% of full-precision accuracy, presenting even better model-size scaling than binary LLMs [20]. This enables dramatic cost reduction and practical deployment on resource-constrained edge devices.

However, traditional (edge-) CPUs face a fundamental architectural misalignment when ternary LLMs are deployed, resulting in degraded performance. State-of-the-art (SOTA) methods such as T-MAC [21] and BitNet.cpp [22] replace multiply-accumulate (MAC) operations with dynamic lookup tables (LUTs) stored in memory (caches and DRAM), shifting workloads from compute-bound to memory-bound ones, achieving  $2.4\text{--}6.2\times$  the throughput compared to the FP16-based kernels [22]. As shown in Fig. 1(c), ternary LUT (TLUT) accesses account for over 75% of system memory requests, creating bandwidth pressure that underutilizes Single-Instruction Multiple-Data (SIMD) execution units ubiquitous in commodity processors [23]. This excessive memory traffic cancels out much of the computational benefit of ternary quantization.

While dedicated accelerators and FPGAs [24]–[26] achieve high ternary inference efficiency, their cost and integration complexity preclude widespread edge deployment. Likewise, server-class features such as Intel AMX [27] or custom ISA extensions with large compute arrays [28]–[31] remain unavailable on edge CPUs due to area and power constraints

[32]. In contrast, the SIMD units that current approaches underutilize already provide high-bandwidth register files with datapaths naturally aligned to ternary operations. Yet no prior work has exploited these SIMD registers for in-register LUT computation, leaving an opportunity to overcome memory bottlenecks and fully harness existing parallel hardware.

Therefore, in this paper we introduce T-SAR, a full-stack co-design framework for scalable, high-throughput ternary LLM inference on edge CPUs, achieved by leveraging existing SIMD hardware with only minimal modifications. **The core innovation of T-SAR is repurposing the SIMD vector register file for dynamic, in-register LUT generation—eliminating costly memory traffic and maximizing data-level parallelism without new compute arrays or complex datapath extensions.** While we focus on the x86 AVX2 ISA, the idea extends naturally to other SIMD ISAs (e.g., ARM NEON, RISC-V Vector [33]), requiring only parameter retuning. T-SAR’s design spans four tightly integrated layers:

- **Algorithmic Layer:** A ternary-to-binary decomposition and data packing scheme for efficient LUT computing.
- **ISA Layer:** Minimal extensions that add register-to-register LUT-based General Matrix Multiplication (GEMM)/General Matrix-Vector Multiplication (GEMV), supporting dynamic computation within SIMD units.
- **Microarchitecture Layer:** Power and area overhead analyses based on lightweight wiring/multiplexing adjustments for SIMD units, validated by ASIC synthesis.
- **Software Layer:** An adaptive kernel dataflow that maximizes throughput across diverse models and platforms.

From high-performance to low-power edge CPUs, T-SAR achieves  $5.6\text{--}24.5\times$  GEMM latency reduction and  $1.1\text{--}86.2\times$  GEMV throughput improvement over SOTA CPU baselines [21], [22] on ternary LLMs from 125M to 100B parameters, with only 3.2% power and 1.4% area overhead in SIMD units. Notably, T-SAR also delivers  $2.5\text{--}4.9\times$  higher energy efficiency than the Jetson AGX Orin GPU on Llama-8B [19] and Falcon3-10B [34]. These results demonstrate that systematic co-design can unlock the latent potential of ubiquitous SIMD hardware for practical edge LLM deployment, narrowing the gap with specialized accelerators while leveraging the world’s most widely deployed compute platform: CPUs.

## II. MOTIVATION

We now characterize the bottleneck in SOTA ternary kernels that motivates the co-design framework T-SAR.

While ternary LLMs promise efficient inference by quantizing weights to  $\{-1, 0, 1\}$  and replacing costly floating-point multiplications with low-cost integer operations, the SOTA approach for executing them introduces a critical bottleneck. As shown in Fig. 2(a,b), these models are implemented in ‘BitLinear’ layers, where the core matrix multiplication is handled by an **LUT-based GEMM/GEMV** method [21], [22]. This technique partitions each input vector (with  $K$  total inputs) into atomic blocks of size  $c$ . For each block, all  $3^c$  possible dot product results are pre-computed and stored in an LUT, transforming the computation from  $O(K)$  arithmetic

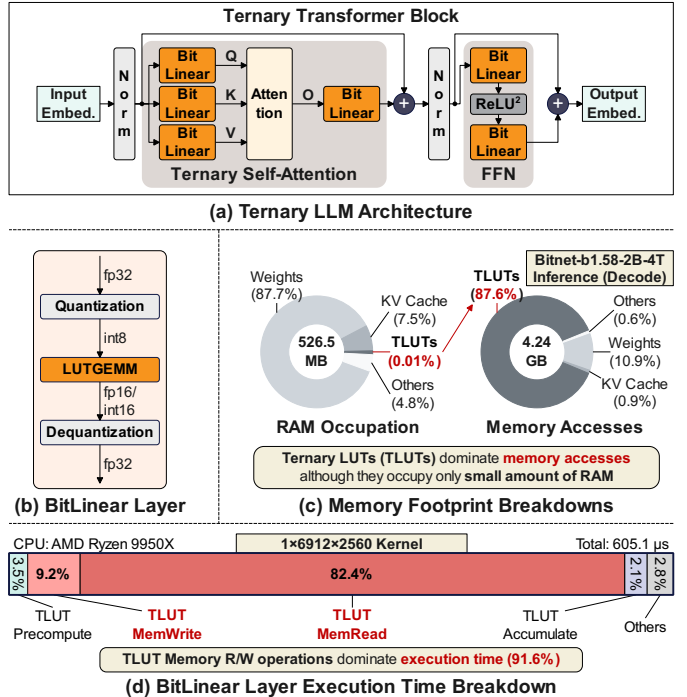


Fig. 2: Ternary LLMs: Architecture and bottleneck analysis. (a) Ternary transformer with BitLinear layers. (b) BitLinear layer workflow including quantization and LUTGEMM. (c) BitNet-b1.58-2B-4T memory footprints: TLUTs, though tiny in RAM, dominate memory accesses. (d) BitLinear GEMV time breakdown: Memory R/W dominates execution.

operations per output to  $O(K/c)$  table lookups. The total LUT storage per layer becomes  $O((K/c) \cdot 3^c)$ , that trades off arithmetic complexity for memory accesses, resulting in the fastest CPU implementations today.

Though theoretically efficient, this trade-off introduces a critical **memory access bottleneck**. As Fig. 3(a) illustrates, the SOTA dataflow relies on pre-computing all LUT values and storing them in system memory. During inference, weights are used to index these LUTs, requiring frequent, random memory accesses that fail to efficiently utilize modern cache hierarchies and powerful SIMD hardware. Our analysis reveals the severity of this issue. Although ternary LUTs (TLUTs) occupy less than 0.01% of total RAM in a representative model, they are accessed so frequently that they account for a staggering **87.6% of all memory transactions** (Fig. 2(c)). This memory saturation means the workload becomes **memory-bound rather than compute-bound**, with **91.6% of the execution time** spent on memory read/write (R/W) operations (Fig. 2(d)).

To overcome these limitations, T-SAR aims to **eliminate external LUT loads and fully exploit the SIMD datapath**, shown in Fig. 3(b). T-SAR generates compressed LUTs on-the-fly inside SIMD registers via custom ISA extensions, eliminating memory TLUT traffic and supporting fused GEMV-accumulation operations to increase throughput. However, achieving in-register LUT computation is non-trivial: the required LUT size ( $3^c$ ) does not match the fixed  $2^n$  bitwidth of SIMD registers, and the limited register file size constrains the number of LUTs that are held simultaneously. This necessitates the careful co-design detailed in the following sections.

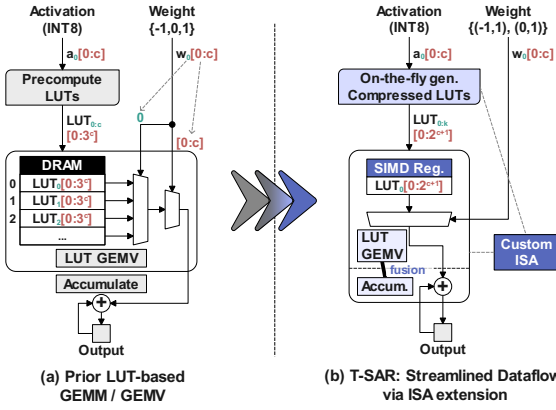


Fig. 3: Prior LUT-based CPU solution vs. T-SAR. (a) Prior: precomputed LUTs loaded from DRAM. (b) T-SAR: on-the-fly compressed LUTs generated in SIMD registers.

### III. T-SAR CO-DESIGN STACK

To resolve the memory bottleneck mentioned in Section II, T-SAR introduces a full-stack co-design that transforms LUT-based operations from memory-bound to compute-bound tasks. The core innovation is that we eliminate memory traffic by generating LUTs on-the-fly, directly within the CPU's high-speed SIMD register file. This section details the tightly integrated algorithmic, ISA, and microarchitectural layers that enable this transformation.

#### A. Algorithm: Ternary-to-Binary Decomposition

The primary challenge of in-register LUT generation is the architectural mismatch between the base-3 nature of ternary weights and the base-2 structure of SIMD hardware. A naive LUT for a block of  $c$  weights would require  $3^c$  entries, which does not align with SIMD register bitwidths.

T-SAR overcomes this with a novel **LUT compression via weight transformation**, as shown in Fig. 4. We decompose a ternary weight block  $\mathbf{w} \in \{-1, 0, 1\}^c$  and its corresponding input activations  $\mathbf{a} \in \mathbb{R}^c$  into two separate binary forms:

- *Dense weights*:  $\mathbf{w}_D \in \{-1, +1\}^c$ , where  $w_{D,i} = w_i$  if  $w_i \neq 0$ , else  $+1$ .
- *Sparse weights*:  $\mathbf{w}_S \in \{0, 1\}^c$ , where  $w_{S,i} = 1$  if  $w_i = 0$ , else  $0$ .

This allows the original dot product to be re-expressed as a subtraction of two binary dot products, removing the influence of the zero-weighted elements:

$$y = \sum_{i=1}^c w_i a_i = \sum_{i=1}^c w_{D,i} a_i - \sum_{i=1}^c w_{S,i} a_i$$

With this transformation, instead of one ternary LUT with  $3^c$  entries, we only need two binary LUTs (for  $\mathbf{w}_D$  and  $\mathbf{w}_S$ ), each of size  $2^c$ . The total storage per block becomes  $2^{c+1}$ , which perfectly matches the power-of-two width of SIMD registers and avoids significant data-path augmentation.

#### B. ISA Extensions for In-Register Execution

As shown in Fig. 5, the decomposition mentioned above enables a two-phase kernel framework. At compile time, ternary weights are encoded into dense and sparse binary

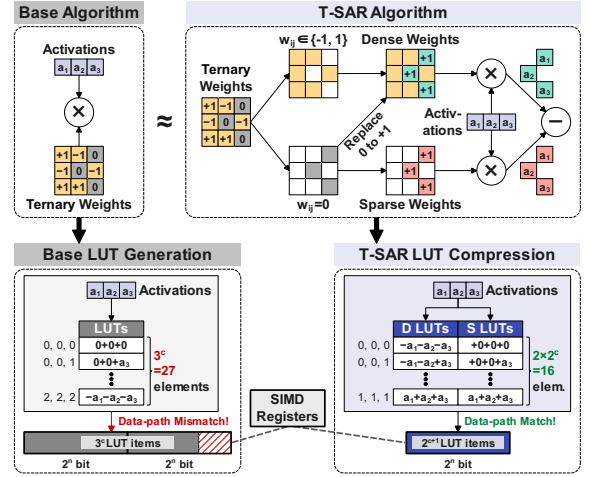


Fig. 4: Proposed LUT GEMV Algorithm for LUT compression, matching the LUT size to the data-path.

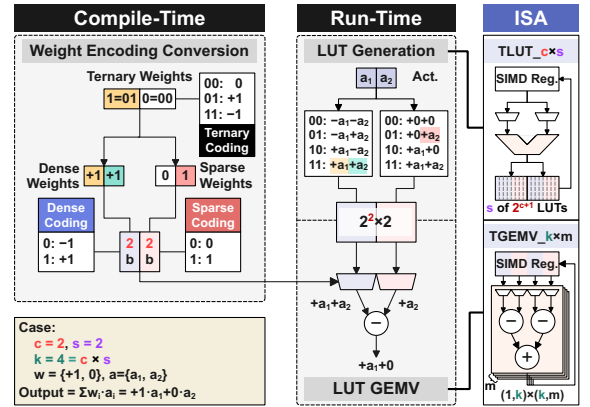


Fig. 5: T-SAR's LUT-based kernel framework overview.

forms. At run time, a minimal set of ISA extensions execute the LUT-based GEMV directly within SIMD registers. These instructions are parameterized by:

- $c$ : the block size.
- $s$ : the number of input blocks processed per instruction.
- $k$ : the number of input channels processed per instruction ( $k = c \times s$ ).
- $m$ : the number of output channels.

The workflow is: (1) the  $\text{TLUT}_{c \times s}$  instruction generates two binary LUTs from activations and places them in SIMD registers; (2) the  $\text{TGEMV}_{k \times m}$  uses these register-resident LUTs with pre-encoded weights to perform the  $(1, k) \times (k, m)$  GEMV; and (3) the final ternary result is reconstructed by subtraction. This register-resident design aligns computation with the SIMD datapath, eliminating the memory bottleneck.

#### C. Microarchitectural Implementation

Fig. 6 demonstrates a practical ISA extension realization on the x86 AVX2 ISA. For the example configuration shown in Fig. 6(a), we set  $c=2$ ,  $s=4$ ,  $k=8$ , and  $m=16$ .

The  $\text{TLUT}_{2 \times 4}$  instruction (Fig. 6(b)) generates four register-resident LUTs, each with  $2^{c+1} = 8$  16-bit entries, occupying two 256-bit YMM registers [35] ( $4 \times 8 \times 16 = 512$  bits). To minimize hardware changes, the operation is split into two  $\mu$ -ops, each writing 256 bits per cycle.

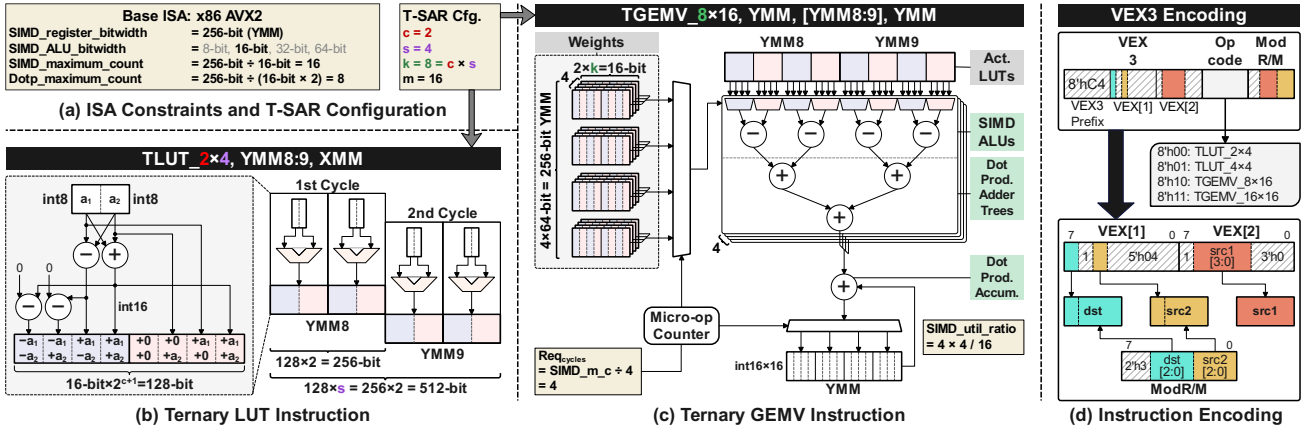


Fig. 6: T-SAR’s ISA extension applied to x86 AVX2 SIMD ISA, demonstrating how the T-SAR instruction primitives are realized with only minimal hardware changes, utilizing existing SIMD ALUs and adder trees. (a) T-SAR configuration and base ISA constraints. (b) TLUT<sub>c</sub>×s example. (c) TGEMV<sub>k</sub>×m example. (d) Instruction encoding details for AVX2, with designed examples of TLUT<sub>c</sub>×s and TGEMV<sub>k</sub>×m instructions.

The TGEMV<sub>8</sub>×16 instruction (Fig. 6(c)) performs a (1, 8)×(8, 16) GEMV, producing 16 outputs. It involves  $s \times m = 64$  subtractions and  $m = 16$  s-to-1 adder tree (ADT) operations, distributed over four  $\mu$ -ops. These instructions reuse the existing 256-bit YMM datapath, including all 16 16-bit SIMD ALUs and 4-to-1 ADTs originally for dot product instructions. Only minor wiring and multiplexer additions are required, keeping area and power overhead minimal.

The instruction encoding, detailed in Fig. 6(d), uses standard VEX3 fields for both TLUT and TGEMV primitives. For instructions that span multiple registers (e.g., TLUT<sub>2</sub>×4 writing to YMM8:9 or TGEMV<sub>8</sub>×16 reading from YMM8:9), the destination or source register is interpreted as a register pair: if dst is 0x1000, the operation uses YMM8 and YMM9.

#### D. Software Kernel and Dataflow Optimization

The performance of GEMM/GEMV operations is highly dependent on data reuse patterns, which vary across different layers of an LLM. To maximize throughput, T-SAR employs an adaptive kernel scheduling strategy. We implement two microkernel dataflows—*activation-persistent* (AP) and *output-persistent* (OP)—to flexibly match these patterns (Fig. 7).

The AP dataflow (Fig. 7(a)) retains input activations in registers across the inner loop, minimizing LUT recomputation and increasing input and weight cache hits. This is effective for layers with high activation and weight reuse (i.e., high  $N$  and  $K$ ). In contrast, the OP dataflow (Fig. 7(b)) keeps output accumulators local until computation completes, reducing memory write-back traffic. This is beneficial in layers with a high number of output channels (high  $M$ ).

At compile-time, T-SAR’s inference framework empirically selects the fastest kernel for each layer, ensuring maximum performance across the entire model.

## IV. EXPERIMENTS AND EVALUATION

We now evaluate T-SAR across diverse models and platforms to validate three key claims: (1) T-SAR delivers significant end-to-end speedups for both prefill (GEMM-heavy) and decode (GEMV-heavy) phases in autoregressive LLMs; (2) these improvements arise from fundamentally reducing the

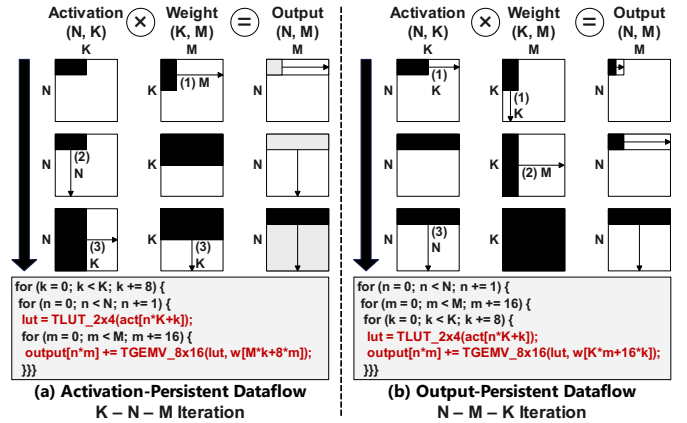


Fig. 7: The T-SAR’s kernel dataflow selections. (a) Activation-persistent dataflow minimizing the TLUT<sub>c</sub>×s invocations and increases input/weight cache hits. (b) Output-persistent dataflow reducing total memory footprints.

memory bottleneck of SOTA LUT-based methods; and (3) they are achieved with minimal hardware overhead, making T-SAR highly efficient even compared to edge GPUs.

#### A. Experimental Setup

**ISA and Simulator:** We extend gem5-AVX 20.1.0.0 [36], [37] (DerivO3CPU) to model the T-SAR ISA. New TLUT<sub>c</sub>×s and TGEMV<sub>k</sub>×m operations are added to the AVX2 pipeline with cycle-accurate  $\mu$ -op sequencing and register-pair reads/writes. Each instruction was verified by executing hand-written assembly with byte-pattern encodings.

**Kernels:** We design three kernel variants for each LUT-GEMV pair (TLUT<sub>2</sub>×4+TGEMV<sub>8</sub>×16, TLUT<sub>4</sub>×4+TGEMV<sub>16</sub>×16), resulting in six kernels: (1) AP-min: activation-persistent with minimal register usage; (2) AP-max: activation-persistent with maximal register use to reduce iterations; (3) OP: output-persistent for minimal write-back traffic. All are implemented in C++ with inline assembly and compiled with GCC 9.4.0.

**Baselines:** We compare against two SOTA LUT-based baselines: Bitnet.cpp TL-2 [22] and T-MAC [21]. To ensure fairness, our kernels include both input quantization and output dequantization stages (shown in Fig. 2(b)).

TABLE I: gem5 simulator configurations for evaluation platforms.

System Type	CPU Model	Simulation Mode	Cores	Freq.	L1 I/D Cache	L2 Cache	L3 Cache	DRAM
<b>Workstation</b>	AMD Ryzen 9950X		16	5.7 GHz	32 KB / 48 KB	1 MB/core	64 MB shared	DDR5-6400 MHz
<b>Laptop</b>	AMD Ryzen 7840U	DerivO3CPU	8	5.1 GHz	32 KB / 32 KB	1 MB/core	16 MB shared	DDR5-4400 MHz
<b>Mobile</b>	Intel Processor N250		4	3.8 GHz	64 KB / 32 KB	2 MB shared	6 MB shared	DDR5-4400 MHz

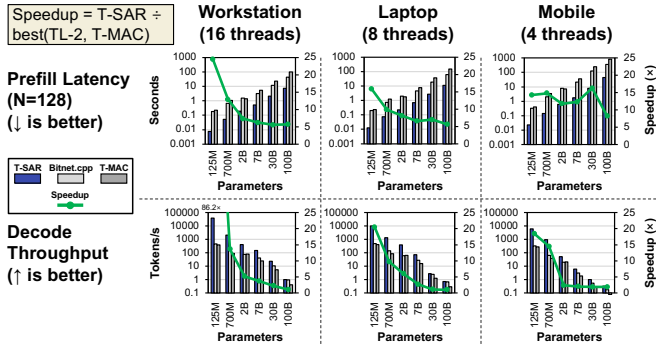


Fig. 8: End-to-end performance across platforms.

**Models and Protocol<sup>1</sup>:** We evaluate BitNet models from 125M to 100B parameters [39]. Prefill runs with  $N = 128$  tokens (batch=1) to build the KV cache; decode measures steady-state throughput using the KV cache. Thread counts are fixed at  $\{16, 8, 4\}$  for  $\{\text{Workstation, Laptop, Mobile}\}$ .

**Metrics:** We report prefill latency, decode throughput, kernel execution time, and kernel memory access requests.

**Platforms:** Three representative x86 CPU classes are modeled (Table I): **Workstation** (Ryzen 9950X, 16 cores), **Laptop** (Ryzen 7840U, 8 cores), **Mobile** (Intel N250, 4 cores).

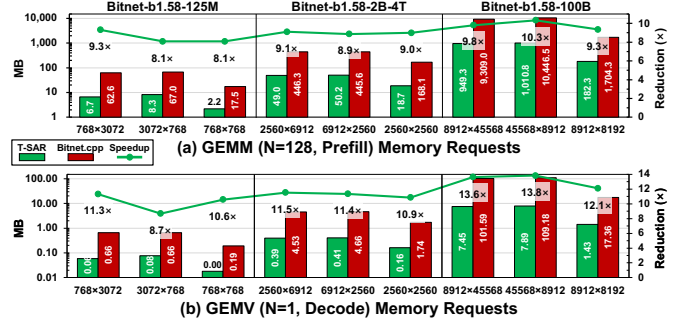
### B. End-to-End Results

**Prefill (GEMM-heavy):** As shown in Fig. 8 (top), T-SAR delivers geo-mean prefill speedups of 8.8 $\times$  (Workstation), 8.4 $\times$  (Laptop), and 12.4 $\times$  (Mobile) across all models (125M–100B). Since GEMM is compute-bound, register-resident LUT generation and fused accumulation let efficiency gains translate almost directly into throughput until cache/memory contention emerges. This explains why prefill benefits exceed decode gains. For example, Mobile’s 7B prefill drops from  $>20$  s to under 1.7 s—enabling interactive LLM use on devices where GPUs cannot be deployed.

**Decode (GEMV-heavy):** Fig. 8 (bottom) shows baseline GEMV is dominated by repeated TLUT fetches. T-SAR removes this traffic entirely, exposing SIMD compute throughput. Relative gains peak on Workstation (6.4 $\times$ ), due to larger caches delaying memory-system bandwidth saturation; Laptop and Mobile reach 4.1–4.2 $\times$ . Mobile’s smaller gain reflects earlier bandwidth saturation despite the request-volume reduction.

**Link to bottlenecks:** The prefill/decode gap mirrors GEMM’s compute-bound and GEMV’s bandwidth-bound nature—setting up the trends seen in memory-system analysis.

<sup>1</sup>Note: We report prefill with  $N=128$  due to simulation cost. TL-2’s weight packing (1.67-bit) is denser than our 1+1-bit split, resulting in approximately 20% more static memory occupation, but end-to-end is dominated by TLUT traffic rather than weight RAM size, hence T-SAR’s advantage. We demonstrate our framework to x86 due to the broad applicability, but retargeting to NEON or RISC-V Vector (RVV) [33] only requires  $c, s, k, m$  tuning due to the different SIMD lane width but extant dot product extensions. For instance, existing ARM NEON’s 128-bit datapath with SDOT/UDOT instruction support (since ARMv8.2-A [38]) realizes the **TLUT\_2 $\times$ 4 + TGEMV\_8 $\times$ 8**.

Fig. 9: Memory request volume (MB) of the kernels executed in the representative BitNet model inference (125M, 2B-4T, 100B). (a) GEMM ( $N=128$ ) prefill. (b) GEMV ( $N=1$ ) decode.

### C. Memory-System Impact

The central claim of T-SAR is that it removes the memory bottleneck by generating LUTs directly in registers. As shown in Fig. 9, this reduces memory request volume (MB) by 8.7–13.8 $\times$  compared to TL-2 [22], with GEMV showing larger relative cuts because the baseline is TLUT-dominated and TLUT fetches are eliminated. For GEMM, TL-2’s denser weight packing (1.67 bits/weight) limits relative reductions, but the resulting stall decreases still improve ALU utilization.

Request reduction grows with model size ( $K \times M$ ) as more LUT calls are avoided, but latency gains diverge: (1) **GEMV case-** Large cuts yield smaller returns as lower Last-level cache (LLC) hit rate reduces effective bandwidth and forces early saturation—most evident in Mobile  $1 \times 8192 \times 45568$ : 89% $\rightarrow$ 62%—capping latency drops. (2) **GEMM case-** Even modest cuts yield large drops since compute-bound phases convert freed cycles and higher locality into utilization; LLC hit rate stays high (89% $\rightarrow$ 91%) until contention. These effects predict the thread-scaling behavior shown in Fig. 10.

### D. Kernel Microbenchmarks and Thread Scaling

From Fig. 10, we make the following observations:

**GEMM case:** For large shapes ( $128 \times 2560 \times 6912$ ,  $128 \times 6912 \times 2560$ ), T-SAR sustains scaling up to 8–16 threads (Workstation) and 4–8 (Laptop) before L3/DRAM contention dominates. Scaling is not perfectly linear, but flattens later than GEMV due to the compute-bound nature and higher data locality, allowing freed cycles from reduced stalls to be fully exploited. This yields up to 13 $\times$  speedup at 4 threads—matching the large prefill gains in Fig. 8.

**GEMV case:** Despite the largest proportional memory savings, GEMV saturates effective bandwidth quickly—often by 2–4 threads on Mobile and 4–8 on Workstation/Laptop—leading to early plateaus and smaller decode-time improvements. Here, extra cores cannot offset bandwidth limits.

Hence, we see that compute-bound kernels (GEMM) sustain scaling and achieve larger absolute gains across platforms, while memory-bound kernels (GEMV) plateau

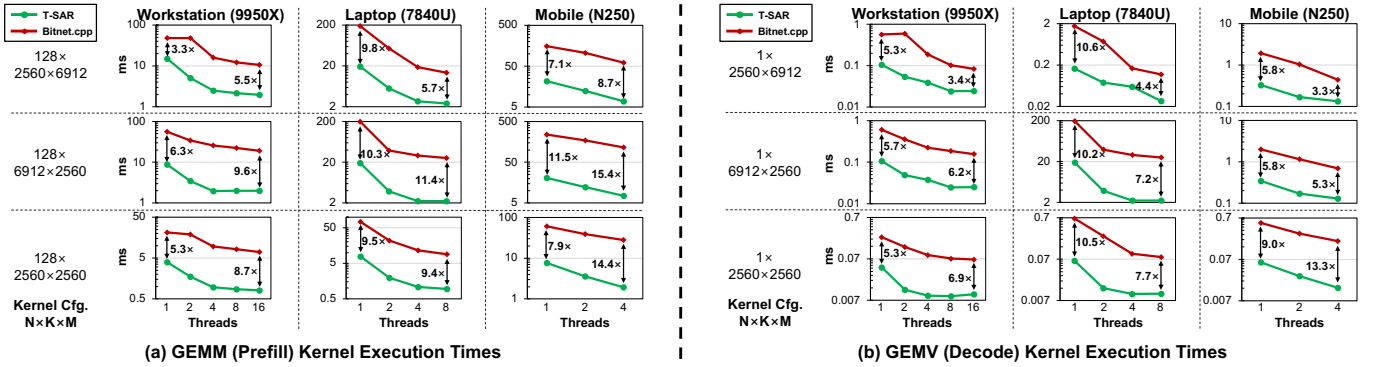


Fig. 10: Multi-thread scaling for BitNet-b1.58-2B-4T. T-SAR vs. TL-2 for GEMM (left) and GEMV (right). Solid lines denote absolute latency (log scale). Arrows show relative speedup.

quickly—highlighting T-SAR’s impact and limits for real-world deployment for edge platforms.

### E. Hardware Overheads

To size the cost of T-SAR, we synthesized a 256-bit SIMD unit (vector add/mul/dot-product, write-back interface) at 1 GHz in TSMC 28 nm using Cadence Genus 21.10, both *without* and *with* the T-SAR logic (shown in Table II). Multi-mode multi-corner (MMMC) synthesis covered  $ssg \leftrightarrow ffg$ ,  $V_{DD} \in [0.81, 1.05]$  V,  $T \in [0, 125]^\circ\text{C}$ ; Area and power are reported at `tt0p9v25c`. The T-SAR instructions reuse existing ALU lanes and register file—no new arithmetic units or scratchpads. Additions are: (i) a 256-bit vector *write-back* MUX to inject TLUT words into the register file, (ii) small operand-bus wires and input MUXes for TLUT/TGEMV (no extra read ports), and (iii) a tiny control/scoreboard block to sequence TLUT writes and fused accumulation.

TABLE II: Synthesis of a 256-bit SIMD slice (TSMC 28 nm, 1 GHz) with and without T-SAR ISA.

Block	Area ( $\mu\text{m}^2$ )			Power (mW)		
	Base	T-SAR	$\Delta$	Base	T-SAR	$\Delta$
SIMD ALUs + write-back interface	73,560	73,560	0.0%	5,904	5,904	0.0%
T-SAR $\rightarrow$ write-back MUX	0	588	<b>+0.8%</b>	0	41	<b>+0.7%</b>
Operand-bus wires and input MUX	0	147	<b>+0.2%</b>	0	24	<b>+0.4%</b>
Others (control/scoreboard, decode)	0	295	<b>+0.4%</b>	0	121	<b>+2.0%</b>
<b>Total</b>	<b>73,560</b>	<b>74,590</b>	<b>+1.4%</b>	<b>5,904</b>	<b>6,090</b>	<b>+3.2%</b>

**Result:** As shown in Table II, the area increases by **+1.4%** ( $73,560 \rightarrow 74,590 \mu\text{m}^2$ ) and active power consumption under kernel-like switching rises by **+3.2%** ( $5,904 \rightarrow 6,090$  mW), dominated by toggling on the new MUX paths. Results correspond to a single 256-bit SIMD slice (no SRAM arrays) and exclude caches and register files; absolute area will scale with the number of slices and integration.

### F. Cross-Platform Comparison

To further evaluate our CPU-based solution, we compare it directly against an edge GPU in the NVIDIA Jetson AGX Orin [40] SoC, using identical model checkpoints and runtime settings (batch=1; steady-state decode). For CPUs, T-SAR power is estimated from CPU *package* power under TL-2 decode via the measured dynamic overhead in Table II:  $P_{\text{T-SAR}} = 1.032 \cdot P_{\text{TL-2}}$ . Energy per token is  $E = P_{\text{T-SAR}} / (\text{tokens/s})$  from the measured T-SAR throughput from gem5 simulator.

TABLE III: Cross-platform decode throughput and energy/token (batch=1). Power boundary: CPU package; GPU module.

Platform	Llama-b1.58-8B		Falcon3-b1.58-10B	
	tokens/s	J/token	tokens/s	J/token
Workstation CPU (9950X, 4nm, T-SAR)	128.96	0.616	103.93	0.795
Laptop CPU (7840U, 4nm, T-SAR)	61.00	0.405	49.65	0.540
Mobile CPU (N250, 10nm, T-SAR)	5.18	0.733	4.30	0.953
Jetson AGX Orin GPU (8nm, llama.cpp)	16.78	1.839	13.25	2.620

**Takeaways.** With matched per-model checkpoints and settings, T-SAR on CPUs outperforms Jetson on **Workstation** and **Laptop** across both families: Llama-b1.58-8B shows 7.7 $\times$  / 3.0 $\times$  (tokens/s / lower J/token) on workstation and 3.6 $\times$  / 4.5 $\times$  on laptop; Falcon3-b1.58-10B shows 7.8 $\times$  / 3.3 $\times$  and 3.7 $\times$  / 4.9 $\times$ , respectively. On **Mobile**, throughput is lower than Jetson (0.31–0.32 $\times$ ), yet energy/token remains 2.5–2.75 $\times$  lower, consistent with our decode memory-bandwidth analysis.

## V. CONCLUSION

We presented T-SAR, a full-stack co-design framework for CPU-only ternary LLM inference. The key idea is to move LUT generation from memory into SIMD registers, turning TLUT fetches into in-register compute and fusing accumulation to shift BitLinear layers from bandwidth-bound to datapath-bound execution. This yields portable speedups across CPUs with no new ALUs and only minor mux/control logic, supported by AP/OP kernels adapt per layer to each platform. Beyond the reported gains, the approach generalizes to RVV and NEON [33] and integrates naturally with sparsity or further quantization. In short, a small ISA extension realigns the bottleneck and enables interactive LLM on CPUs, even mobile ones, while keeping hardware changes minimal.

## ACKNOWLEDGEMENTS

This work was supported in part by the DARPA Young Faculty Award, the National Science Foundation (NSF) under Grants #2127780, #2319198, #2321840, #2312517, and #2235472, #2431561, the Semiconductor Research Corporation (SRC), the Office of Naval Research through the Young Investigator Program Award, and Grants #N00014-21-1-2225 and #N00014-22-1-2067, Army Research Office Grant #W911NF2410360. Additionally, support was provided by the Air Force Office of Scientific Research under Award #FA9550-22-1-0253, along with generous gifts from Xilinx and Cisco.

## REFERENCES

- [1] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.
- [2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288*, 2023.
- [3] Y. Zheng, Y. Chen, B. Qian, X. Shi, Y. Shu, and J. Chen, "A review on edge large language models: Design, execution, and applications," *ACM Comput. Surv.*, vol. 57, no. 8, Mar. 2025. [Online]. Available: <https://doi.org/10.1145/3719664>
- [4] K. Ahi, "Risks & Benefits of LLMs & GenAI for Platform Integrity, Healthcare Diagnostics, Cybersecurity, Privacy & AI Safety: A Comprehensive Survey, Roadmap & Implementation Blueprint," *arXiv preprint arXiv:2506.12088*, 2025.
- [5] X. Ma, G. Fang, and X. Wang, "LLM-Pruner: On the Structural Pruning of Large Language Models," in *Adv. Neural Inf. Process. Syst. 36 (NeurIPS)*, New Orleans, LA, USA, 2023, pp. 21 702–21 720.
- [6] Q. Fu, M. Cho, T. Merth, S. Mehta, M. Rastegari, and M. Najibi, "LazyLLM: Dynamic Token Pruning for Efficient Long Context LLM Inference," *arXiv preprint arXiv:2407.14057*, 2024.
- [7] H. Chen, Y. Ni, W. Huang, S. Yang, H. Oh, Y. Liu, T. Das, and M. Imani, "LVLM\_CSP: Accelerating Large Vision Language Models via Clustering, Scattering, and Pruning for Reasoning Segmentation," in *ACM Int. Conf. Multimed. (MM)*, Dublin, Ireland, 2025, p. 3932–3941.
- [8] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, "AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration," in *Annu. Conf. Mach. Learn. Syst. 6 (MLSys)*, vol. 6, Santa Clara, CA, USA, 2024, pp. 87–100.
- [9] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "GPT3.int8(): 8-Bit Matrix Multiplication for Transformers at Scale," in *Adv. Neural Inf. Process. Syst. 35 (NeurIPS)*, New Orleans, LA, USA, 2022, pp. 30 318–30 332.
- [10] Y. Xu, X. Han, Z. Yang, S. Wang, Q. Zhu, Z. Liu, W. Liu, and W. Che, "Onebit: Towards extremely low-bit large language models," in *Adv. Neural Inf. Process. Syst. 37 (NeurIPS)*, Vancouver, BC, Canada, Dec. 2024, pp. 66 357–66 382.
- [11] P. Palacios, R. Medina, G. Ansaloni, and D. Atienza, "HEEPstor: an Open-Hardware Co-design Framework for Quantized Machine Learning at the Edge," in *22nd ACM Int. Conf. Comput. Front.: Workshops Spec. Sess. (CF)*, Cagliari, Italy, 2025, p. 22–25. [Online]. Available: <https://doi.org/10.1145/3706594.3726967>
- [12] A. Kundu, F. Lim, A. Chew, L. Wynter, P. Chong, and R. D. Lee, "Efficiently Distilling LLMs for Edge Applications," *arXiv preprint arXiv:2404.01353*, 2024.
- [13] Y. Ma, C. Shen, L. Jiang, T. Xu, and M. Zhang, "TKD: An Efficient Deep Learning Compiler with Cross-Device Knowledge Distillation," in *Des. Autom. Test Eur. (DATE)*, Lyon, France, 2025, pp. 1–7.
- [14] S. Jeong, H. E. Barkam, H. Oh, H. Chen, T. Das, Z. Ye, and M. Imani, "iTaskSense: Task-Oriented Object Detection in Resource-Constrained Environments," in *62nd ACM/IEEE Des. Autom. Conf. (DAC)*, San Francisco, CA, USA, 2025, pp. 1–7.
- [15] H. Bai, W. Zhang, L. Hou, L. Shang, J. Jin, X. Jiang, Q. Liu, M. Lyu, and I. King, "BinaryBERT: Pushing the Limit of BERT Quantization," in *Jt. Conf. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. (ACL-IJCNLP)*, Virtual, Aug. 2021, pp. 4334–4348.
- [16] H. Qin, Y. Ding, M. Zhang, Q. YAN, A. Liu, Q. Dang, Z. Liu, and X. Liu, "BiBERT: Accurate Fully Binarized BERT," in *Int. Conf. Learn. Represent. (ICLR)*, Virtual, Apr. 2022.
- [17] Z. Yuan, Y. Shang, and Z. Dong, "PB-LLM: Partially binarized large language models," in *Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024.
- [18] H. Wang, S. Ma, L. Dong, S. Huang, H. Wang, L. Ma, F. Yang, R. Wang, Y. Wu, and F. Wei, "BitNet: Scaling 1-bit Transformers for Large Language Models," *arXiv preprint arXiv:2310.11453*, 2023.
- [19] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, "The Era of 1-Bit LLMs: All Large Language Models Are in 1.58 Bits," *arXiv preprint arXiv:2402.17764*, 2024.
- [20] A. Kaushal, T. Vaidhya, A. K. Mondal, T. Pandey, A. Bhagat, and I. Rish, "Surprising Effectiveness of Pretraining Ternary Language Model at Scale," in *Int. Conf. Learn. Represent. (ICLR)*, 2025, pp. 1–48.
- [21] J. Wei, S. Cao, T. Cao, L. Ma, L. Wang, Y. Zhang, and M. Yang, "T-MAC: CPU Renaissance via Table Lookup for Low-Bit LLM Deployment on Edge," in *21st European Conf. on Computer Syst. (EuroSys)*, Rotterdam, Netherlands, Mar. 2025, pp. 278–292.
- [22] J. Wang, H. Zhou, T. Song, S. Cao, Y. Xia, T. Cao, J. Wei, S. Ma, H. Wang, and F. Wei, "Bitnet.cpp: Efficient Edge Inference for Ternary LLMs," *arXiv preprint arXiv:2502.11880*, 2025.
- [23] D. Vuță-Popescu, I. C. Antofi, C. B. Ciobanu, and C. Z. Kertész, "Simd extensions—a historical perspective," in *2024 IEEE 30th International Symposium for Design and Technology in Electronic Packaging (SI-ITME)*. IEEE, 2024, pp. 108–115.
- [24] C. Yin, Z. Bai, P. Venkatram, S. Aggarwal, Z. Li, and T. Mitra, "TerEffic: Highly Efficient Ternary LLM Inference on FPGA," *arXiv preprint arXiv:2502.16473*, 2025.
- [25] Y. Qiao, Z. Chen, Y. Zhang, Y. Wang, and S. Huang, "TeLLMe: An Energy-Efficient Ternary LLM Accelerator for Prefilling and Decoding on Edge FPGAs," *arXiv preprint arXiv:2504.16266*, 2025.
- [26] R. Chen, J. Liu, S. Tang, Y. Liu, Y. Zhu, M. Ling, and B. Da Silva, "ATE-GCN: An FPGA-Based Graph Convolutional Network Accelerator with Asymmetrical Ternary Quantization," in *Des. Autom. Test Eur. (DATE)*, Lyon, France, 2025, pp. 1–6.
- [27] S. Na, G. Jeong, B. H. Ahn, A. Jezghani, J. Young, C. J. Hughes, T. Krishna, and H. Kim, "FlexInfer: Flexible LLM Inference with CPU Computations," in *Annu. Conf. Mach. Learn. Syst. 7 (MLSys)*, 2025.
- [28] G. Armeniakos, A. Maras, S. Xydias, and D. Soudris, "Mixed-Precision Neural Networks on RISC-V Cores: ISA Extensions for Multi-Pumped Soft SIMD Operations," in *IEEE/ACM Int. Conf. Comput. Aided Des. (ICCAD)*, Newark, NJ, USA, Oct. 2024, pp. 1–9.
- [29] A. Garofalo, G. Tagliavini, F. Conti, D. Rossi, and L. Benini, "XpulpNN: Accelerating Quantized Neural Networks on RISC-V Processors Through ISA Extensions," in *Des. Autom. Test Eur. (DATE)*, Virtual, Mar. 2020, pp. 186–191.
- [30] N. K. Purayil, M. Perotti, F. Fischer, and L. Benini, "AraXL: A Physically Scalable, Ultra-Wide RISC-V Vector Processor Design for Fast and Efficient Computation on Long Vectors," in *Des. Autom. Test Eur. (DATE)*, Lyon, France, 2025, pp. 1–7.
- [31] H. W. Oh and S. E. Lee, "The Design of Optimized RISC Processor for Edge Artificial Intelligence Based on Custom Instruction Set Extension," *IEEE Access*, vol. 11, pp. 49 409–49 421, 2023.
- [32] H. W. Oh, S. An, W. S. Jeong, and S. E. Lee, "RF2P: A Lightweight RISC Processor Optimized for Rapid Migration from IEEE-754 to Posit," in *IEEE/ACM Int. Symp. Low Power Electron. Des. (ISLPED)*, Vienna, Austria, 2023, pp. 1–6.
- [33] J.-H. Li, J.-K. Lin, Y.-C. Su, C.-W. Chu, L.-T. Kuok, H.-M. Lai, C.-L. Lee, and J.-K. Lee, "SIMD Everywhere Optimization from ARM NEON to RISC-V Vector Extensions," *arXiv preprint arXiv:2309.16509*, 2023.
- [34] Falcon-LLM Team, "The Falcon 3 Family of Open Models," Dec. 2024, available: <https://huggingface.co/blog/falcon3>.
- [35] C. Lomont, "Introduction to intel advanced vector extensions," *Intel white paper*, vol. 23, no. 23, pp. 1–21, 2011.
- [36] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 Simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, p. 1–7, Aug. 2011. [Online]. Available: <https://doi.org/10.1145/2024716.2024718>
- [37] S. Lee, Y. Kim, D. Nam, and J. Kim, "Gem5-AVX: Extension of the Gem5 Simulator to Support AVX Instruction Sets," *IEEE Access*, vol. 12, pp. 20 767–20 778, 2024.
- [38] Arm Ltd., "A64 SIMD Vector Instructions," <https://developer.arm.com/documentation/100069/0609/A64-SIMD-Vector-Instructions>, accessed: March 26, 2025.
- [39] J. Wang, H. Zhou, T. Song, S. Mao, S. Ma, H. Wang, Y. Xia, and F. Wei, "1-bit AI Infra: Part 1.1, Fast and Lossless BitNet b1.58 Inference on CPUs," *arXiv preprint arXiv:2410.16144*, 2024.
- [40] M. Barnell, C. Raymond, S. Smiley, D. Isereau, and D. Brown, "Ultra Low-Power Deep Learning Applications at the Edge with Jetson Orin AGX Hardware," in *IEEE High Perform. Extreme Comput. Conf. (HPEC)*, 2022, pp. 1–4.