SCARLET: A Scalable OPCM-Based Accelerator for Transformer Inference with Tiled Crossbars

Sina Karimi*, Guowei Yang*, Carlos A. Ríos Ocampo†, Ajay Joshi*, Ayse K. Coskun*
*Boston University, †University of Maryland, College Park
{sikarimi, guoweiy, joshi, acoskun}@bu.edu, riosc@umd.edu

Abstract—**While transformer-based large language models (LLMs) have achieved state-of-the-art performance on a wide range of natural language processing tasks, their massive computational demands, especially during inference, pose a significant challenge. Photonic accelerators offer a promising solution, but existing designs struggle with the precision, dynamism, and storage requirements of modern LLMs. This paper introduces SCARLET, a hybrid photonic architecture that addresses these limitations through two key components. First, we design a high-density optical phase-change memory (OPCM) crossbar for static matrix multiplications, achieving 5.6× higher bit density and 86.43% lower energy compared to previous OPCM crossbar designs. Second, we introduce an approximate photonic floating-point multiplier to handle dynamic matrix multiplications and quantization steps by approximating floating-point computations with weighted integer sums, thus, eliminating the need for frequent memory reprogramming. Our evaluation on models with up to 13 billion parameters demonstrates significant performance improvements, including up to 17.15× and 8.45× lower latency during prefill and generation phases, respectively.**

Index Terms—**artificial intelligence, silicon photonics, hardware accelerators, large language models, Phase change memory**

I. Introduction

Transformer-based large language models (LLMs) have significantly advanced state-of-the-art natural language and image processing [1], [2]. However, widely-used GPUs are inefficient for multi-billion-parameter inference, primarily due to the memory-bound generation stage and multi-headed attention overhead [3]. While prior work has proposed various digital CMOS accelerator architectures (e.g., [4], [5]), incorporating both hardware specialization and algorithmic optimizations, these designs now face growing power and scalability challenges. This is because the end of Dennard scaling has led to increased power density without proportional performance gains [6].

Integrated photonic accelerators have emerged as a compelling alternative for next-generation computing platforms [7], offering ultra-high-speed performance and low energy consumption. Various optical systems are being explored, including Mach-Zehnder interferometer (MZI) arrays [8], [9], micro-ring resonator banks [10], and PCM-based crossbars [11]. However, transformers introduce unique challenges. First, their self-attention mechanism uses dynamic, input-dependent matrices, making frequent reprogramming prohibitively expensive for architectures with long reprogramming time such as PCMs [12], [8]. Second, mitigating accuracy loss in large LLMs requires mixed-precision computation to avoid errors from full quantization [13]. Current photonic solutions are limited to fixed-point values and cannot transition between precision domains, forcing costly data movement to a host for conversions that undermines their performance and efficiency.

To address these shortcomings, we propose **SCARLET**, a Scalable Crossbar-based Accelerator for Reduced-precision LLM Execution using Tiled Photonic Arrays. Our design combines two key components: First, we design a high-density Optical Phase-Change Memory (OPCM) crossbar tailored for static matrix multiplications in decoder layers, achieving 5.6× higher bit density and reduced energy consumption compared to prior designs. Second, SCARLET includes a new optical processing unit capable of performing element-wise multiplications in full precision for both quantization/dequantization steps and attention score/output calculations. This unit approximates floating-point multiplications using weighted integer sums [14], enabling efficient dynamic computation directly in the photonic domain with minimal accuracy loss.

By co-optimizing these components, our hybrid architecture efficiently handles both static and dynamic computations in the LLM decoder, while also supporting seamless precision changes between stages. This design significantly improves scalability, energy efficiency, and performance. We focus on the decoder layer because it is the primary bottleneck for inference latency and energy consumption in LLMs. The decoder generates new tokens sequentially, a process that is highly demanding and critical to overall system performance. Our contributions are as follows:

- We design SCARLET, a hybrid photonic architecture for LLM inference that uses high-density OPCM crossbars for static multiplications and photonic crossbars with approximate floating-point units for dynamic attention and quantization/dequantization. Our OPCM crossbar achieves 5.6× higher bit density, lowering photonic loss and reprogramming overhead in static operations.
- We optimize the quantization pipeline by fusing dequantization with dynamic computations, cutting conversion overhead and reducing energy by up to 69.2% compared to electronic designs.
- Evaluations on OPT, Llama-2, and GPTNeo models with 2.7B up to 13B parameters show up to 17.15× latency reduction in decoder prefill and 8.45× in generation compared to NVidia L40S GPUs.

II. Background

A. OPCM Background

SCARLET consists of multiple OPCM chiplets, each containing 2D arrays of $Ge_2Sb_2Te_5$ (GST)-based cells integrated into optical waveguides (Figure 1). GST is a phase-change material that can reversibly switch between amorphous and crystalline states using optical pulses or Joule heating [15]. Intermediate states, achieved by partial crystallization, yield distinct optical transmittance levels and enable multi-bit, non-volatile storage without standby energy [16].

Switching can be performed electrically via microheaters or optically with pump pulses. Graphene microheaters have demonstrated efficient electrical switching, consuming 5.55 nJ for amorphization and 860.71 pJ for crystallization [12]. Optical switching is more energy efficient at the cell level (460 pJ / 140 pJ) [16], but scalability is limited due to the need for complex routing or large-area grating couplers [17]. Thus, electrical switching remains the more practical option for large-scale and scalable integration [18].

Recent work shows up to 64 stable states per GST cell (6 bits), enabling analog in-memory computation [19], [20]. Scalar multiplication is achieved by encoding the multiplicand in GST transmittance and the multiplier in the amplitude of an incident pulse, producing an output proportional to their product. Arranging such cells in crossbar arrays naturally extends this to matrix–vector multiplications, the core operation in our design [17].

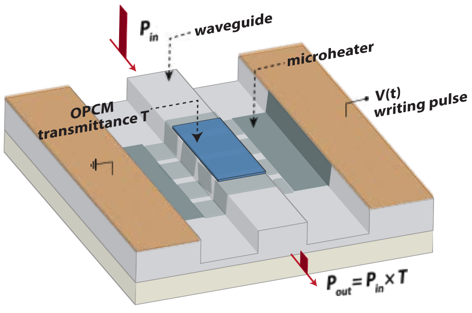

Fig. 1. OPCM device: an integrated multi-bridge microheater [21] configures the OPCM (a GST cell) via Joule heating. The OPCM changes its transmittance (T), modulating the amplitude of a pulse propagating through the waveguide.

B. Transformer model Background

The Transformer architecture [1] revolutionized NLP by replacing sequential models such as RNNs, offering strong parallelization and the ability to capture long-range dependencies through self-attention [22]. While the original model used an encoder-decoder structure, LLMs for generative tasks typically adopt a decoder-only variant.

Self-attention enables the model to weigh token relationships using query (Q), key (K), and value (V) matrices:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

with multi-head attention extending this across multiple representation subspaces. Each token is then processed by a feed-forward network (FFN):

$$\text{FFN}(x) = ACT(xW_1)W_2 \qquad (2)$$

where ACT is typically ReLU, GELU, or Swish.

Input text is tokenized into subwords (e.g., BPE, WordPiece), mapped to IDs, and embedded into dense vectors before being processed. A key difference from traditional networks is that Transformers perform dynamic matrix multiplications. Q, K, and V are generated at inference time, making attention input-dependent and also more computationally demanding for hardware acceleration.

C. Quantization

In SCARLET and prior photonic designs [8], [23], matrix multiplications are restricted to fixed-point precision, requiring quantization of both weights and activations. Quantization reduces memory and compute by mapping values to low-bit integers (e.g., INT8) [24], but retaining activations in floating point is impractical for photonic hardware [25]. In this work, we adopt absolute maximum (absmax) quantization, which scales all values by the largest absolute value in the tensor or vector.

Fully quantizing large LLMs is difficult due to the wide dynamic range of activations. Naive uniform quantization often causes accuracy loss [13], [24], so finer granularities are used. Per-tensor quantization applies a single scale factor, while per-token or per-channel quantization assigns local scales to capture variability at higher metadata cost. SmoothQuant [13] addresses imbalance by rescaling inputs and weights, enabling accurate post-training INT8 quantization without retraining activations in full precision.

Certain operations, such as layer normalization and softmax, remain in higher precision to avoid quality degradation, while the rest of the model benefits from quantization's efficiency.

III. SCARLET

In this section, we present SCARLET, a mixed-precision architecture for efficient end-to-end LLM decoder inference. SCARLET adopts a hybrid strategy: static weight multiplications are mapped to high-density OPCM arrays, while dynamic multiplications and quantization/dequantization steps are executed on a separate photonic crossbar using approximate floating-point operations [8], [23]. This partitioning eliminates costly OPCM reprogramming for dynamic data and enables a full decoder layer with optimized dataflow between quantized and full-precision domains. Prior work mitigates quantization challenges by keeping sensitive operations in high precision [13]. We extend this mixed-precision approach with a hardware-centric design that executes dynamic multiplications in full precision, notably the attention computations over Q, K, and V. Our proposed execution flow shown in figure 2 minimizes precision-conversion overhead by dequantizing Q, K, and V immediately after their computation, and re-quantizing only after the attention output.

In SCARLET, dequantization and element-wise multiplication are fused in a single photonic-crossbar cycle, avoiding intermediate transfers and reducing Analog-to-Digital Converter

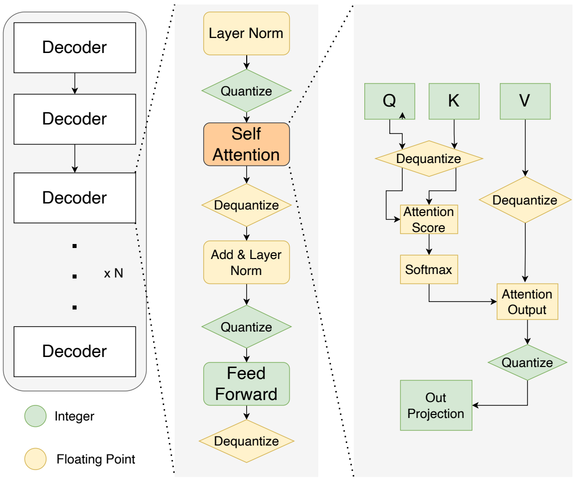

Fig. 2. Execution flow of the decoder layer in SCARLET. All dynamic matrix multiplication operations are preceded by a dequantization step, allowing to perform both scaler multiplication for dequantization and element-wise multiplication for matrix multiplication in one cycle in quant/dequant unit.

(ADC) cost while improving energy efficiency in attention score and output computation.

A. System architecture

Figure 3 illustrates the high-level architecture of SCARLET, a 2.5D-integrated system for LLM decoder acceleration. It comprises a host, main memory, and the accelerator integrating distinct chiplets: OPCM, quant/dequant, High Bandwidth Memory (HBM), controller, and digital units.

The controller manages communication between the host and chiplets, executing pre-generated schedules with a simple state machine. The HBM3 chiplet stores layer weights, KV-cache, and intermediate buffers. OPCM chiplets handle static INT8 multiplications such as Q, K, V projections and FFN blocks. The quant/dequant chiplet performs precision conversion, dynamic multiplications for attention scores and outputs in floating point, and accumulation via FP adders. Finally, the digital units chiplet executes non-linear functions including layernorm, softmax, ReLU, and GeLU.

B. Microarchitecture

1) OPCM Chiplet: Each OPCM chiplet contains multiple PEs, where each PE consists of an OPCM crossbar, SRAM buffer, control logic, and programming circuitry. Weights for static matrix multiplications are programmed into the GST cells of the crossbar. Unlike prior designs that required separate arrays for positive and negative values [26], our design encodes both within a single array using four GST cells per intersection (24 bits total), reducing area footprint and enabling direct representation of 8-bit values by splitting positive and negative components across paired cells.

During inference, input rows are modulated onto laser signals via electro-optical converters and passed through the crossbar in two cycles (positive and negative values). GST cells attenuate the signals to perform analog multiplication, while photodetectors accumulate results in the electrical domain, reducing optical loss compared to prior all-photonic accumulation [11], [26]. Dedicated waveguides connect each GST cell, mitigating insertion and splitting losses of directional couplers and enabling higher bit density and parallelism.

Our architecture computes complete K and V matrices directly on OPCM arrays, while Q is incrementally generated for efficient streaming into the Quant/Dequant chiplet, reducing transfer delay and improving attention throughput.

2) Quant/dequant chiplet: To handle precision conversion and dynamic matrix multiplications, SCARLET integrates a dedicated Quant/Dequant chiplet, avoiding the high reprogramming overhead of OPCM arrays.

The chiplet uses a hardware-efficient approximation of floating-point multiplication [14], where operands are represented as integers by concatenating exponent and mantissa, and multiplication is approximated with integer addition plus a fixed bias. Implemented with photonic crossbars that only perform additions, this method supports both quantization scaling and element-wise multiplications without requiring multipliers, with negligible accuracy loss.

The design includes one main array along with multiple secondary arrays interconnected to the main array (2×256 and 4×256) matched to OPCM throughput (Section III-B1). Rows supply scalers for quantization/dequantization and bias for error correction, while outputs from the main array feed the secondary arrays for element-wise multiplication with scaling. Column outputs are collected by photodetectors and ADCs, with the inter-array connection reducing conversion bottlenecks.

Finally, dequantization is fused with dynamic multiplications (QK^T and AV), so inputs go through precision converters, are modulated optically, and combined with bias in the crossbar, amortizing ADC/DAC cost and improving efficiency.

3) Electrical Chiplet: Certain operations, such as softmax and layer normalization, are not well-suited for OPCM-based computation due to their complexity and precision requirements. Previous works have implemented these functionalities using standard digital circuits, which provide the necessary flexibility and efficiency. Our design includes dedicated digital units optimized for speed and accuracy, complementing the analog computation in OPCM arrays. This split between photonic and digital domains is a common strategy in prior architectures [8], [27], enabling efficient handling of both linear and non-linear operations.

C. Dataflow

The execution flow of SCARLET partitions static and dynamic operations across the OPCM, quant/dequant, and digital chiplets, with HBM serving as the global memory. This co-design ensures that weights are programmed once, inputs are streamed with low overhead, and intermediate results are routed to the next unit.

1) Static Operations in OPCM: Prior to inference, the OPCM arrays are programmed with the static weights of the decoder layer, including projection matrices and feed-forward network (FFN) weights. Using the multi-cell encoding scheme described in Section III-B1, each tile of the weight matrix is mapped to a dedicated array, and the programmed state remains

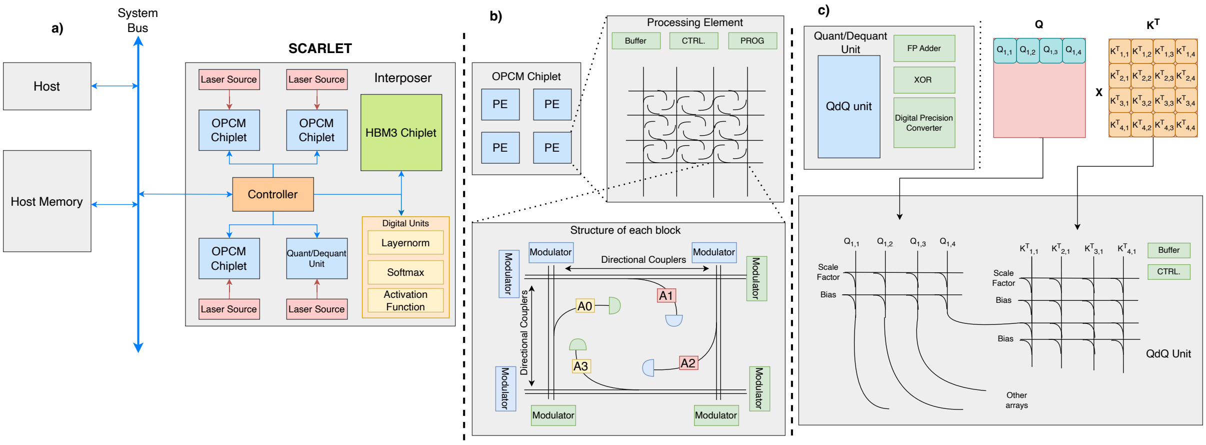

Fig. 3. a) SCARLET is a 2.5D accelerator connected to the host via a system bus. All data, including weights and inputs, are transferred to the HBM chiplet on the accelerator. b) Each OPCM chiplet contains multiple PE units, and each PE integrates an OPCM crossbar array together with buffers, control logic, and reprogramming circuitry. Each array cell is composed of four GST cells, and parallel waveguides run along the array, with each waveguide connected to the GST cells along its path. c) The Quant/Dequant unit consists of one 2×256 array connected to 256 4×256 arrays. During multiplication, the row of the first matrix is injected into the initial array, while the columns of the second matrix are injected into the remaining arrays.

fixed throughout execution. During inference, input activations are retrieved from HBM and converted into optical signals by electro-optical (E–O) modulators. These signals propagate through the OPCM arrays, where GST cells attenuate the light according to the stored weights, performing analog multiplications. Parallel waveguides allow multiple inputs to be injected simultaneously, each distributed to the GST cells connected along its path. At each intersection, the attenuated signals are collected by photodetectors, which convert the output pulse into electrical current and accumulate results in the electrical domain. This approach reduces optical loss compared to prior all-photonic accumulation methods, while enabling subtraction by combining outputs from positive and negative weight cells. The resulting partial sums are then either written back to HBM or forwarded directly to the next compute stage.

2) Dynamic Operations in Quant/Dequant: The Quant/Dequant chiplet processes outputs from the OPCM arrays when precision conversion or dynamic computation is required. For quantization and dequantization, scaling factors are applied using approximate floating-point multiplication implemented with integer additions in photonic crossbars. This approach avoids costly floating-point multipliers while retaining sufficient accuracy for attention computations. In addition, the chiplet executes dynamic matrix multiplications such as QK^T and AV, which require operands generated at runtime. Performing these matrix multiplications in OPCM units incurs significant delay due to reprogramming latency. These operations are fused with dequantization, so scaling and multiplication are performed together in a single photonic-crossbar cycle. This design reduces data movement and amortizes the energy cost of ADC/DAC conversions, which are major bottlenecks.

3) Non-linear Operations in Digital Units: The results of matrix multiplications and quantization stages are finally sent to the digital chiplet, which handles operations unsuited to photonic arrays, including layer normalization, softmax, and activation functions such as ReLU and GeLU. These functions are implemented in dedicated digital units optimized for speed and accuracy, ensuring robust execution of precision-sensitive components while maintaining system throughput.

In summary, the dataflow in SCARLET follows a structured path: static weights are encoded in OPCM arrays for efficient INT8 multiplications, intermediate activations are routed to the Quant/Dequant chiplet for scaling and dynamic floating-point operations, and nonlinear transformations are carried out in digital units. By carefully partitioning computation across chiplets and minimizing intermediate conversions, the architecture balances the efficiency of analog photonic computing with the flexibility of digital logic, enabling scalable execution of full decoder layers.

IV. Methodology and Evaluation

A. Evaluation Methodology

To assess the impact of our proposed techniques on model quality, we build on the original Hugging Face implementations and use their FP16 models as the baseline. We apply our modifications—including quantization, approximate multiplication, and SmoothQuant optimization [13]—and measure accuracy against the FP16 baseline. All accuracy and performance experiments are conducted on NVIDIA L40S GPUs [28], using token-level next-token prediction accuracy. The evaluation included OPT (2.7B, 6.7B, 13B), Llama-2 (7B, 13B), and GPTNeo (2.7B) models.

At the system level, each accelerator consists of a controller, an HBM3 memory chiplet, three OPCM chiplets for static matrix multiplications, a Quant/Dequant chiplet for precision conversion and dynamic operations, a digital chiplet for non-linear functions, and a laser source. Each OPCM chiplet occupies 456.95 mm^2 and integrates 11 PEs, each a 256×256 OPCM array with 4.2 MB of SRAM. Based on directional couplers and waveguide dimensions [17], each OPCM cell measures 25×25 μm^2. The Quant/Dequant chiplet occupies 108.8 mm^2 and includes 257 arrays, with a total of 51.2 MB

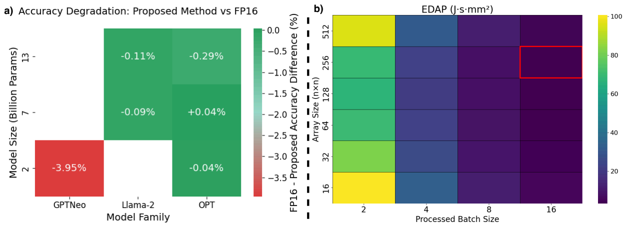

Fig. 4. a) Accuracy degradation due to the proposed approximations and quantization. OPT and GPTNeo models use a single scaling factor per tensor, while Llama-2 models use per-token/per-channel quantization. b) Energy–Delay–Area Product (EDAP) of a single accelerator executing the FFN layer of OPT-13B with varying array sizes and batch sizes. The configuration with a 256×256 array and a batch size of 16 achieves the optimal EDAP.

SRAM and FP adders to match throughput. Because these arrays do not use GST cells, their size is determined by waveguide arcs and couplers at 15 × 15 μm^2. The digital chiplet implements layer normalization [29], softmax [30], and activation functions for the FFN block [31].

We conservatively set the accelerator's operating frequency to 5 GHz, below the 18 GHz demonstrated for OPCM devices [17], to reflect practical peripheral circuit constraints. The total power includes both optical and electrical contributions. The laser's power is the primary component of the total optical power consumption. Laser power is derived from optical loss (0.6 dB GST cell, 0.0028 dB crossing, 0.01 dB coupler [17]) and combined quantum efficiency of 10%. This results in 128 mW per wavelength. PCM programming consumes 5.55 nJ (amorphization) and 860.71 pJ (crystallization) with a 400 ns reprogramming latency [12]. Electro-optical conversion costs 1 pJ/bit [17], while O–E conversion at 5 GS/s consumes 29 mW [32]. On-chip SRAM, generated using GF22FDX and scaled to 7 nm [33], [34], consumes 1.39 W. Floating-point units (e.g., precision converters, multipliers, adders) are synthesized using the Berkeley HardFloat library [35]. HBM accesses incur 20 pJ/bit [36].

The accelerator relies on a 16-lane CXL interface with 64 GB/s aggregate bandwidth. HBM latency is modeled as 40 ns for accesses within the same interposer and 80 ns across interposers [37]. The global controller orchestrates inter-chiplet data movement, coordinating transfers between HBM and the local SRAM buffers of the OPCM, quant/dequant, and digital chiplets. Control logic is kept minimal, relying on pre-generated schedules. Together, these parameters define the baseline for our system-level performance and efficiency evaluation.

B. Accuracy Evaluation

To study the impact of quantization and approximation on model accuracy, we modify Hugging Face implementations of the target models and compare them against their FP16 baselines. Decoder layers are replaced with our custom implementation that integrates INT8 quantization, approximate floating-point multiplication, and SmoothQuant optimization [13]. We evaluate accuracy on the WikiText benchmark [38]. Our modifications replace multiplications in attention scores, attention outputs, and quantization/dequantization scaling with the approximate method [14]. OPT and GPT models use full-tensor quantization, while Llama models apply per-channel or per-token quantization.

As shown in Figure 4, accuracy loss is at most 0.29% for OPT and Llama-2, due to SmoothQuant scaling and the low error of approximate multiplication. GPTNeo suffers a larger 3.95% drop under per-tensor quantization; accuracy can be improved with per-token or per-channel scaling, at the cost of additional overhead. For Llama, per-tensor quantization causes significant degradation, which we address by applying per-token or per-channel quantization to activations and weights.

C. Microarchitecture comparison

1) OPCM Chiplet: Varying the size of OPCM crossbars impacts area, laser power, buffer size, ADC count, and overall performance. Larger arrays improve efficiency by amortizing reprogramming but increase SRAM buffer requirements, creating a trade-off between performance and memory. Figure 4 reports EDAP for the FFN layer of OPT-13B (512 tokens), a compute-intensive decoder stage. The best EDAP is achieved with a 256×256 array and batch size of 16, selected as the practical upper bound for LLM inference workloads.

We also compare our OPCM crossbar against the prior design from [17] using a 256×256 configuration running the FFN layer of OPT-13B across different sequence lengths. As shown in Figure 5, for sequence length 512 our design achieves 10.96× lower delay and 86.43% energy savings stem from reduced photonic losses via parallel waveguides, eliminating extra couplers and cutting laser power by three orders of magnitude per cycle. Higher parallelism enables multiple inputs per cycle, reducing inference time, while increased bit density lowers reprogramming overhead. Since weight matrices remain constant across sequence lengths, reprogramming energy and delay are identical for both designs.

2) Quant/dequant unit: To evaluate the Quant/Dequant unit, we compare a photonic implementation against an electronic counterpart synthesized with the Berkeley HardFloat library [35]. The comparison measures dequantization energy during attention score computation for sequence length 512 and batch size 16, with both designs matched to OPCM throughput.

The photonic design contains 257 arrays—one for the input row and 256 for matrix columns. As shown in Figure 6, it achieves substantially lower energy than the electronic baseline across all models. These savings arise from performing dequantization and multiplication in the optical domain, amortizing

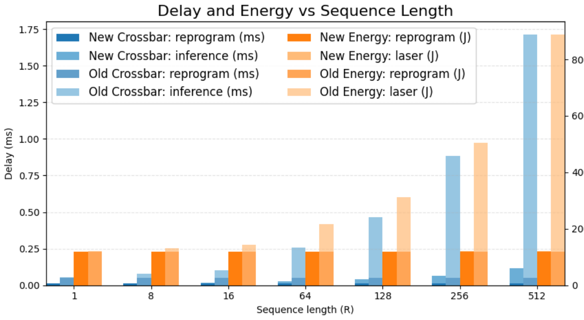

Fig. 5. Comparison of the proposed and previous OPCM crossbar designs when executing the FFN layer of OPT-13B with a batch size of 16 for different sequence lengths in delay and energy. Our design achieves 10.96× reduction in delay and 86.43% lower energy consumption for sequence length of 512.

ADC costs across row/column operations. Accumulation occurs later in the electronic domain, further improving efficiency.

These results highlight the significant energy reduction achievable for attention score computation, demonstrating the potential of photonic systems for efficient LLM inference.

D. Architecture comparison

To evaluate performance, we compare SCARLET's decoder against the FP16 baseline running on GPU, measuring latency in both the prefill (processing input tokens) and generation (token-by-token output) phases. Because our design specifically targets large models with complex quantization execution flows on photonic accelerators, a direct comparison with prior work is infeasible. To the best of our knowledge, this is the first framework to address the specific challenges of deploying large-scale quantized models on analog accelerators. All models use the same input of 210 tokens with batch size 16, and SCARLET is configured with 256×256 arrays (Section III-B1). Figure 7 reports average decoder latency for SCARLET and GPU runs with operation breakdowns.

In the prefill stage, SCARLET achieved speedups of 1.16×–3.31× for OPT and Llama-2, and 17.22× for GPTNeo. Prefill requires more computation, leading to higher delay in the quantization and OPCM units. In the generation stage, speedups were larger: 3.34×–8.11×, with GPTNeo reaching 8.61×. Here, data movement dominated latency, especially for models with complex quantization. OPT and GPTNeo use simple tensor-level quantization with minimal overhead, while Llama-2 employs per-row and per-column scaling, adding significant quantization delay. Overall, GPTNeo showed poorer GPU performance, consistent with prior findings [39], which resulted in more significant speedup in our design.

V. Related Work

Prior work on accelerating Transformers using photonic accelerators has largely focused on smaller models such as DistilBERT and BERT. Zhu et al. [8] proposed directional-coupler cells with balanced photodetectors, which support full-range values but lack weight storage and require modulation of both inputs; the design is also an order of magnitude larger than OPCM cells. Afifi et al. [23] explored MRR banks to store weights with low compute energy, but these require continuous refresh and lack the passive storage of PCM-based designs. Shen et al. [9] implemented MZI arrays, which allow full-range modulation but are slow to program and area-intensive. These photonic designs have not been scaled beyond 6.7B parameters, where quantization error becomes critical, motivating the need for dedicated quantization/dequantization units and algorithmic modifications.

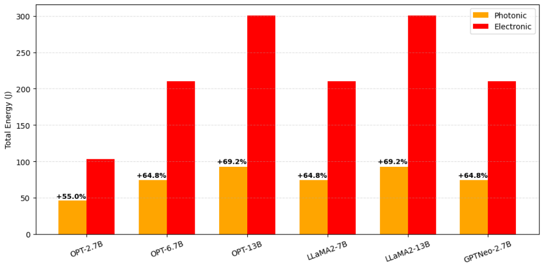

Fig. 6. Energy consumption comparison between photonic and electronic dequantizer units for attention score calculation across various models. Both designs are configured to match the throughput of OPCM units.

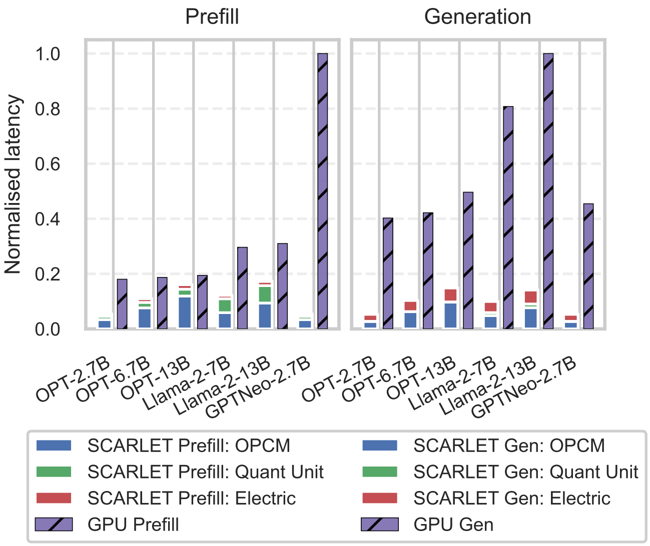

Fig. 7. This comparison shows the normalized average decoder layer latency for SCARLET and a GPU during a 210-token prefill and a subsequent generation phase, using an identical KV-cache size. The memory-bound generation stage has a significantly higher data transfer delay compared to the prefill stage. Furthermore, Llama-2-13B's more complex quantization (per-token/per-column) demands more operations from the quant unit.

Electrical approaches have also been investigated. Park et al. [5] accelerate attention with a PIM-based unit while executing the compute-intensive layers on a GPU, reducing memory bottlenecks. Seo et al. [4] proposed Ianus, a unified memory architecture that integrates an NPU with a PIM engine, coordinating access to balance compute- and memory-bound tasks. Although effective at smaller scales, electrical designs face scalability challenges, as the end of Dennard scaling has led to rising power densities without proportional performance gains [6].

VI. Conclusion

We presented SCARLET, a hybrid photonic accelerator that uses OPCM crossbars for static multiplications and a photonic Quant/Dequant unit for dynamic operations. SCARLET maintains accuracy loss compared to FP16 baselines, reduces OPCM delay and energy by up to 10.96× and 86.4%, and outperforms GPUs with up to 17.2× speedup in prefill and 8.6× in generation. These results highlight the potential of photonic-electronic architectures for scalable and energy-efficient LLM inference.

Acknowledgment

This research was partially funded by the NSF CCF 2131127 grant.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient streaming language models with attention sinks," *arXiv preprint arXiv:2309.17453*, 2023.

[3] S. Hong, S. Moon, J. Kim, S. Lee, M. Kim, D. Lee, and J.-Y. Kim, "Dfx: A low-latency multi-fpga appliance for accelerating transformer-based text generation," in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2022, pp. 616–630.

[4] M. Seo, X. T. Nguyen, S. J. Hwang, Y. Kwon, G. Kim, C. Park, I. Kim, J. Park, J. Kim, W. Shin *et al.*, "Ianus: Integrated accelerator based on npu-pim unified memory system," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, 2024, pp. 545–560.

[5] J. Park, J. Choi, K. Kyung, M. J. Kim, Y. Kwon, N. S. Kim, and J. H. Ahn, "Attacc! unleashing the power of pim for batched transformer-based generative model inference," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2024, pp. 103–119.

[6] M. M. Waldrop, "The chips are down for moore's law," *Nature News*, vol. 530, no. 7589, p. 144, 2016.

[7] J. Gu, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "Light in ai: toward efficient neurocomputing with optical neural networks—a tutorial," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 6, pp. 2581–2585, 2022.

[8] H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan, "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2024, pp. 686–703.

[9] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature photonics*, vol. 11, no. 7, pp. 441–446, 2017.

[10] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "Crosslight: A cross-layer optimized silicon photonic neural network accelerator," in *2021 58th ACM/IEEE design automation conference (DAC)*. IEEE, 2021, pp. 1069–1074.

[11] G. Yang, C. Demirkiran, Z. E. Kizilates, C. A. R. Ocampo, A. K. Coskun, and A. Joshi, "Processing-in-memory using optically-addressed phase change memory," in *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2023, pp. 1–6.

[12] Z. Fang, R. Chen, J. Zheng, A. I. Khan, K. M. Neilson, S. J. Geiger, D. M. Callahan, M. G. Moebius, A. Saxena, M. E. Chen *et al.*, "Ultra-low-energy programmable non-volatile silicon photonics based on phase-change materials with graphene heaters," *Nature Nanotechnology*, vol. 17, no. 8, pp. 842–848, 2022.

[13] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "Smoothquant: Accurate and efficient post-training quantization for large language models," in *International conference on machine learning*. PMLR, 2023, pp. 38 087–38 099.

[14] H. Luo and W. Sun, "Addition is all you need for energy-efficient language models," *arXiv preprint arXiv:2410.00907*, 2024.

[15] J. R. Erickson, N. A. Nobile, D. Vaz, G. Vinod, C. A. Ríos Ocampo, Y. Zhang, J. Hu, S. A. Vitale, F. Xiong, and N. Youngblood, "Comparing the thermal performance and endurance of resistive and pin silicon microheaters for phase-change photonic applications," *Optical Materials Express*, vol. 13, no. 6, pp. 1677–1688, 2023.

[16] C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "Integrated all-photonic non-volatile multi-level memory," *Nature photonics*, vol. 9, no. 11, pp. 725–732, 2015.

[17] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[18] R. Chen, Z. Fang, F. Miller, H. Rarick, J. E. Froch, and A. Majumdar, "Opportunities and challenges for large-scale phase-change material integrated electro-photonics," *ACS Photonics*, vol. 9, no. 10, pp. 3181–3195, 2022.

[19] C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, "Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network," *Nature communications*, vol. 12, no. 1, p. 96, 2021.

[20] C. Ríos, N. Youngblood, Z. Cheng, M. Le Gallo, W. H. Pernice, C. D. Wright, A. Sebastian, and H. Bhaskaran, "In-memory computing on a photonic platform," *Science advances*, vol. 5, no. 2, p. eaau5759, 2019.

[21] H. Sun, C. Lian, F. Vásquez-Aza, S. Rahimi Kari, Y.-S. Huang, A. Restelli, S. A. Vitale, I. Takeuchi, J. Hu, N. Youngblood *et al.*, "Microheater hotspot engineering for spatially resolved and repeatable multi-level switching in foundry-processed phase change silicon photonics," *Nature Communications*, vol. 16, no. 1, p. 4291, 2025.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] S. Afifi, F. Sunny, M. Nikdast, and S. Pasricha, "Tron: Transformer neural network acceleration with non-coherent silicon photonics," in *Proceedings of the great lakes symposium on VLSI 2023*, 2023, pp. 15–21.

[24] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale," *Advances in neural information processing systems*, vol. 35, pp. 30 318–30 332, 2022.

[25] A. Sobhanan, A. Fardoost, D. Desai, F. G. Vanani, Z. Zhu, S. S. Pang, and G. Li, "Photonic floating point multiplication using cascaded ssb-sc modulation," *Optics Express*, vol. 32, no. 22, pp. 39 177–39 191, 2024.

[26] G. Yang, S. Karimi, C. A. R. Ocampo, A. K. Coskun, and A. Joshi, "Sophie: A scalable recurrent ising machine using optically addressed phase change memory," in *2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2024, pp. 1548–1561.

[27] J. Park, J. Choi, K. Kyung, M. J. Kim, Y. Kwon, N. S. Kim, and J. H. Ahn, "Attacc! unleashing the power of pim for batched transformer-based generative model inference," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2024, pp. 103–119.

[28] NVIDIA Corporation, "NVIDIA L40S GPU Specifications," https://www.nvidia.com/en-us/data-center/l40s/, 2025, accessed July 2025.

[29] H. Guo, L. Peng, J. Zhang, Q. Chen, and T. D. LeCompte, "Att: A fault-tolerant reram accelerator for attention-based neural networks," in *2020 IEEE 38th International Conference on Computer Design (ICCD)*. IEEE, 2020, pp. 213–221.

[30] M. Wang, S. Lu, D. Zhu, J. Lin, and Z. Wang, "A high-speed and low-complexity architecture for softmax function in deep learning," in *2018 IEEE asia pacific conference on circuits and systems (APCCAS)*. IEEE, 2018, pp. 223–226.

[31] H. Prashanth and M. Rao, "Somalib: Library of exact and approximate activation functions for hardware-efficient neural network accelerators," in *2022 IEEE 40th International Conference on Computer Design (ICCD)*. IEEE, 2022, pp. 746–753.

[32] M. Guo, J. Mao, S.-W. Sin, H. Wei, and R. P. Martins, "A 5 gs/s 29 mw interleaved sar adc with 48.5 db sndr using digital-mixing background timing-skew calibration for direct sampling applications," *IEEE Access*, vol. 8, pp. 138 944–138 954, 2020.

[33] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost, M. Buehler, V. Chikarmane, T. Ghani, T. Glassman *et al.*, "A 22nm high performance and low-power cmos technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density mim capacitors," in *2012 symposium on VLSI technology (VLSIT)*. IEEE, 2012, pp. 131–132.

[34] A. Shafaei, Y. Wang, X. Lin, and M. Pedram, "Fincacti: Architectural analysis and modeling of caches with deeply-scaled finfet devices," in *2014 IEEE Computer Society Annual Symposium on VLSI*. IEEE, 2014, pp. 290–295.

[35] U. B. P. C. Laboratory, "berkeley-hardfloat: Parameterized floating-point units in chisel," https://github.com/ucb-bar/berkeley-hardfloat, 2025, accessed July 2025.

[36] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*. IEEE, 2014, pp. 10–14.

[37] A. Cho, A. Saxena, M. Qureshi, and A. Daglis, "A case for cxl-centric server processors," *arXiv preprint arXiv:2305.05033*, 2023.

[38] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," *arXiv preprint arXiv:1609.07843*, 2016.

[39] P. J. Maliakel, S. Ilager, and I. Brandic, "Investigating energy efficiency and performance trade-offs in llm inference across tasks and dvfs settings," *arXiv preprint arXiv:2501.08219*, 2025.