A Light-Speed Large Language Model Accelerator with Optical Stochastic Computing

Salma Afifi
Colorado State University
Fort Collins, USA
Salma.Afifi@colostate.edu

Oluwaseun Alo
University Of Kentucky
Lexington, USA
seun.alo@uky.edu

Ishan Thakkar
University of Kentucky
Lexington, USA
igthakkar@uky.edu

Sudeep Pasricha
Colorado State University
Fort Collins, USA
sudeep@colostate.edu

Abstract

To address the increasingly intensive computational demands of attention-based large language models (LLMs), there is a growing interest in developing energy-efficient and high-speed hardware accelerators. To that end, photonics is being considered as an alternative technology to digital electronics. This work introduces a novel optical hardware accelerator that leverages stochastic computing principles for LLMs. Our proposed accelerator incorporates full-range optical stochastic multipliers and stochastic-analog compute-capable optical-to-electrical transducer units to efficiently handle static and dynamic tensor computations in attention-based models. Our analysis shows that our accelerator exhibits at least 7.6× speedup and 1.3× lower energy compared to state-of-the-art LLMs hardware accelerators.

CCS Concepts

• **Optical computing**; • **Application-specific VLSI designs**; • **Neural Networks**;

Keywords

Transformer neural networks, silicon photonics, inference acceleration, stochastic computing, optical computing

ACM Reference Format:
Salma Afifi, Oluwaseun Alo, Ishan Thakkar, and Sudeep Pasricha. 2025. A Light-Speed Large Language Model Accelerator with Optical Stochastic Computing. In *Great Lakes Symposium on VLSI 2025 (GLSVLSI '25), June 30–July 02, 2025, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3716368.3735299

1 Introduction

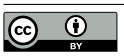
Large language models (LLMs) have become foundational to modern natural language processing (NLP) and computer vision, powering breakthroughs in machine translation, question answering, and image recognition. Leveraging powerful attention mechanisms, LLMs excel at capturing complex, long-range dependencies within data, producing highly accurate and context-aware outputs [1]. Pioneering models such as Transformer [1], BERT [2], GPT [3], and Vision Transformers (ViTs) [4] have set new benchmarks across a range of applications. However, this impressive accuracy comes at a significant computational cost: with billions of parameters, LLMs impose substantial inference latencies, energy consumption, and processing complexity.

To mitigate these challenges, increasing computational demands have spurred the development of specialized hardware accelerators aimed at improving inference efficiency. While several digital accelerators have incorporated LLM-specific optimizations to improve performance [5],[6], traditional platforms increasingly face limitations due to the breakdown of Dennard scaling, leading to higher power densities and diminishing performance gains [7]. Thus, researchers are increasingly turning to alternative computing paradigms such as silicon photonics, which offers light-speed data transmission, high parallelism, and improved energy efficiency.

Silicon photonic accelerators have demonstrated promise in accelerating neural network operations, with successful deployments across various deep neural networks (DNNs) [7]-[12]. However, these accelerators face key challenges, including costly and frequent electro-optic conversions and inter-channel or heterodyne crosstalk. Moreover, most are optimized for weight-static architectures like convolutional neural networks (CNNs), making them less suitable for the dynamic workloads of LLMs.

To the best of our knowledge, this paper presents the first optical accelerator that leverages stochastic computing for LLMs. Through incorporating stochastic and analog computing principles, our accelerator efficiently exploits the high-bandwidth and energy efficiency gains inherent to optical computing while mitigating several of its challenges. The novel contributions of this paper are:

- We propose a novel silicon photonic-based hardware accelerator that integrates stochastic and analog computing for LLMs.
- We develop a novel optical vector dot-product (VDP) core featuring homodyne optical VDP elements (VDPEs) that eliminate the reliance on high-cost digital-to-analog (DAC) devices, mitigate heterodyne crosstalk, significantly reduce insertion loss, and lower overall laser power requirements.
- We design a dynamically operated optical stochastic signed multiplier (OSSM) that enables highly parallel and energy-efficient dynamic full-range matrix multiplications.



This work is licensed under a Creative Commons Attribution 4.0 International License.

GLSVLSI '25, New Orleans, LA, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1496-2/2025/06
https://doi.org/10.1145/3716368.3735299

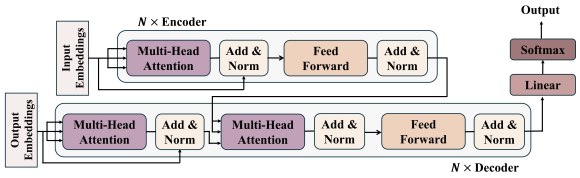


Figure 1: Transformer neural network architecture overview.

- We propose several device-, architecture-, and circuit-level optimizations to support low-latency optical computations.
- We perform a comprehensive comparison with GPU, TPU, CPU, and several state-of-the-art accelerators for LLMs.

2 Background

2.1 LLM Workload Models

LLMs have become foundational in machine learning, particularly for tasks with long-term dependencies. Their attention mechanism enables efficient handling of long-range relationships. Most LLMs are built on the transformer architecture [1], which comprises encoder and decoder blocks: the encoder transforms input sequences into continuous, high-dimensional representations, while the decoder generates output tokens based on encoded data and prior outputs. Each block includes two main components: multi-head self-attention (MHA) and feed-forward network (FFN), as shown in Figure 1.

The MHA is composed of H number of heads. In each head, the input is transformed into query (Q), key (K), and value (V) matrices by linear projection. These matrices are used to compute the attention scores via a scaled dot-product operation as:

$$Head\,(I) = attention\,(Q,k,V) = softmax\left(QK^T/\sqrt{d_k}\right).V \quad (1)$$

where d_k is the dimension of Q and K. The attention mechanism generates its output by concatenating the results from multiple attention heads, followed by a linear transformation. The FFN typically comprises two dense layers with an activation function—such as GELU or ReLU—applied in between. Modern transformer-based models like BERT [2] use a stack of N encoder blocks, followed by a feed-forward layer and GELU activation. Similarly, ViT [4] employs N encoder layers followed by a multi-layer perceptron. In contrast, models such as GPT-4 [3] rely solely on decoder blocks. While LLMs have achieved remarkable success, their implementation on hardware accelerators, particularly photonic-based ones, presents notable challenges. Unlike the static weight matrices in linear layers and traditional neural networks such as CNNs, the attention mechanism's dynamic nature—requiring the generation of Q, K, and V matrices at runtime—introduces significant complexities for acceleration. These difficulties are further exacerbated by factors like crosstalk noise in optical systems, making the efficient optical hardware acceleration of LLMs a highly challenging endeavor.

2.2 Stochastic Computing

Stochastic computing (SC) reduces computational complexity by representing values with sequences of individual bits, trading precision for simpler logic design and lower power consumption. Due to these efficiencies, SC has gained traction in areas such as signal processing, control systems, and DNNs [13], [14]. Stochastic computing represents real numbers through probabilistic bit-streams, where the occurrence rate of 1s to 0s reflects the real value. Eq. 2) and (3) show examples of stochastic number representation:

$$X_1 = \frac{6}{10} \rightarrow x_1\,(stoch.) = 0110101101 \quad (2)$$

$$X_2 = \frac{4}{10} \rightarrow x_2\,(stoch.) = 1010010001 \quad (3)$$

After this encoding step, computations are performed by statistically manipulating the input bit-streams, allowing many standard binary functions to be performed with simple logic gates rather than complex circuits [13]-[17]. For instance, a multiplication operation in SC can be executed using a single AND gate on two stochastic bitstreams. Multiplying the numbers from Eq. 2) and (3) would be computed as:

$$X_1 \times X_2 = x_1 \& x_2 = 0010000001\ (= 0.2) \quad (4)$$

Note that the product of X_1 and X_2 is expected to yield a real value of 0.24, yet the bitwise AND operation of x_1 and x_2 produces a result of 0.2, illustrating potential precision loss in SC. Our accelerator introduces specialized methods to overcome such inaccuracies and enhance computational precision.

2.3 Optical Analog ANN Acceleration

Optical ANN accelerators have garnered strong interest from both academia and industry for their high performance and energy efficiency [7]-[12], [18]-[20]. These architectures typically operate in either a coherent or non-coherent manner. Coherent designs encode parameters in the optical signal's phase to perform multiply-accumulate (MAC) operations [12], while non-coherent designs modulate the signal's amplitude. Parallelism is achieved using multi-wavelength signals and banks of opto-electric modulators, often based on microring resonators (MRs). MRs manipulate the optical amplitude to perform computations and are tuned to specific resonant wavelengths (λ_{MR}), defined as:

$$\lambda_{MR} = \frac{2\pi R}{m} n_{eff} \quad (5)$$

where R is the MR radius, m is the order of the resonance, and n_{eff} is the effective index of the device. Electronic data can be modulated onto the optical signal passing an MR by carefully adjusting n_{eff} (and hence λ_{MR}) with a tuning circuit. Figure 2 illustrates an optical non-coherent VDP core, and also shows an MR modulator in the activation bank (corresponding to λ_{MR}) imprinting an activation value onto the signal transmission.

Non-coherent optical accelerators usually leverage wavelength-division multiplexing (WDM) to boost throughput by combining multiple signals into a single waveguide [12], as shown in Figure 2. A set of N distinct wavelengths, generated by laser diodes, is multiplexed into one waveguide and split into M branches using a 1×M splitter. Each branch contains a VDPE that performs N multiplications using two MR bank arrays: one encodes activations, the other weights. A balanced photodetector (BPD) aggregates outputs from the positive and negative weight arms into an analog signal, which is then digitized and summed in a reduction unit.

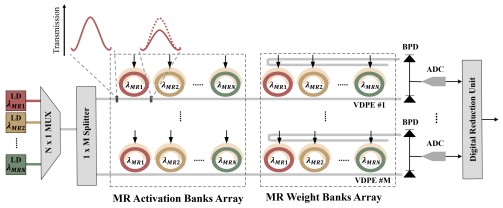


Figure 2: An optical VDP core.

While conventional architectures offer substantial parallelism as demonstrated in [7]-[9], several key limitations hinder their scalability and full exploitation of optical bandwidth. First, insertion losses can be significant, as each optical signal per VDPE experiences losses from 2 in-band MRs and 2×(N-1) out-of-band MRs. Second, heterodyne crosstalk, caused by interference between different wavelengths, is a major noise source in WDM systems. Device-level analysis in [9] shows that maintaining 8-bit precision limits usable wavelengths to 18 (or 36 MRs per VDPE), constraining scalability. Third, reliance on DACs for encoding model parameters into analog signals for MR tuning adds cost and complexity. Lastly, static weight mapping to MR banks enforces weight-stationary (WS) dataflows, limiting support for more flexible schemes like output-stationary (OS). While these issues are manageable for CNN, RNN, and GNN accelerators, they present significant challenges for LLMs.

TRON [11], a recent non-coherent optical accelerator for transformer models, incorporates architecture-level optimizations to reduce opto-electronic conversions during inference. However, its throughput remains constrained by the challenges outlined earlier and requires doubling the number of VDPEs to support signed values. Similarly, Lightning-Transformer [10] introduces a photonic tensor core using a crossbar array of interference-based optical VDPEs with MZMs. While effective, the design requires additional MZMs to encode sign bits via phase modulation, and DAC power consumption remains a key performance bottleneck.

In contrast to prior solutions, we introduce a novel optical computing substrate built on OSSMs. Each VDP core is divided into wavelength-specific VDPEs, with each VDPE operating on a single wavelength, eliminating heterodyne crosstalk. OSSMs encode single-bit stochastic values (ON/OFF power levels) onto optical signals, simplifying multiplications to bitwise AND to enable massive parallelism and drastically reduce power consumption and optical dynamic range compared to multi-level signals in non-coherent analog accelerators. The stochastic format also eliminates the need for DACs.

3 Proposed Hardware Accelerator

In contrast to prior solutions, we introduce a novel optical computing substrate built on OSSMs. Each VDP core is divided into wavelength-specific VDPEs, with each VDPE operating on a single wavelength, eliminating heterodyne crosstalk. OSSMs encode single-bit stochastic values (ON/OFF power levels) onto optical signals, simplifying multiplications to bitwise AND to enable massive parallelism and drastically reduce power consumption and optical dynamic range compared to multi-level signals in non-coherent analog accelerators. The stochastic format also eliminates the need for DACs.

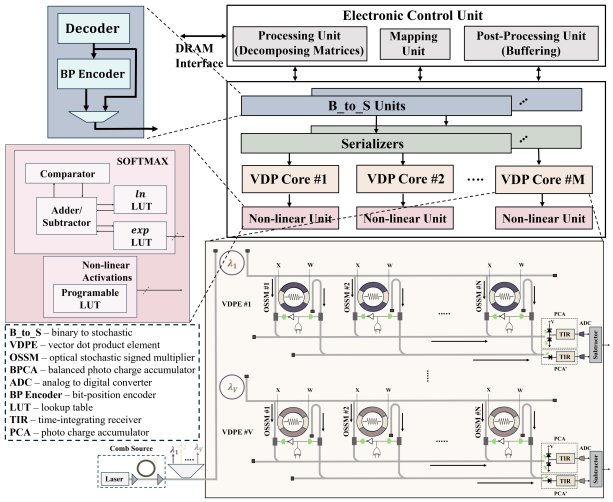


Figure 3: ASTRA architecture overview.

This section presents an overview of ASTRA, our optical LLM accelerator, with its architecture illustrated in Figure 3. The main computational tasks—general matrix multiplications (GEMM)—are performed within the VDP cores. Stochastic operands are generated via binary-to-stochastic (B_to_S) converters and high-speed serializers. An integrated electronic control unit (ECU) handles main memory interfacing, operand buffering, resource allocation, and weight matrix mapping. The following subsections detail ASTRA's architecture and implemented optimizations.

3.1 VDP Cores

The VDP cores in ASTRA employ a novel architectural design, departing from the conventional non-coherent WDM based approach discussed in Section 2.3. A low-power laser comb source generates V wavelengths that are coupled into a waveguide, and routed by MRs into distinct VDPEs, each composed of N OSSMs. All multiplication operations are performed stochastically through logical AND operations as described in Section 2.2. To maintain accuracy and avoid the degradation typically associated with stochastic platforms [14]-[17], we employ a low-cost and low-error deterministic B_to_S conversion method [16]. This technique encodes the first operand using a binary-to-transition-coded-unary (B_to_TCU) decoder followed by a bit-position correlation encoder, while the second operand is encoded using only a B_to_TCU decoder, as shown in the B_to_S unit in Figure 3. TCU numbers are stochastic bit-vectors where all the '1's are grouped at either of the stream's trailing ends. The two stochastic bit-vectors are serialized and multiplied using the OSSM units that reside in VDP cores. The homodyne optical output streams (i.e., carried by the same wavelength) from all OSSMs within a VDPE are routed through separate lanes for positive and negative values. Photodetectors (PDs) perform incoherent superposition and charge accumulation [12], converting these optical signals into analog electrical pulses. The resulting pulses are added to previously stored charges in the

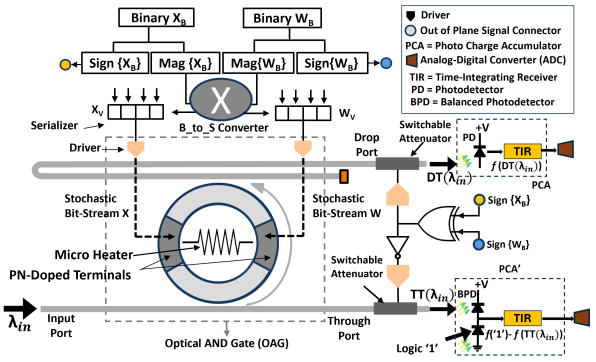


Figure 4: Schematic of our OSSM.

photo-charge accumulators (PCA for positive, PCA' for negative). Finally, ADCs digitize the outputs, and a digital subtractor computes the final result.

The VDP core achieves high efficiency by leveraging optical computing while addressing limitations of conventional designs [14]-[17]. ASTRA integrates S_to_B and optical-to-electrical conversions, and accumulation within the PCA and ADC structures. Heterodyne crosstalk is eliminated by assigning a single wavelength per VDPE, and coherent crosstalk is avoided due to the system's incoherent operation. Optical insertion losses are also reduced, as each signal traverses only 2 in-band MRs, 1 filter MR and 1 OSSM, compared to 2 in-band and $2 \times (N - -1)$ out-of-band MRs in conventional designs (see Section 2.3).

3.2 Optical Stochastic Signed Multiplier (OSSM)

Our OSSM, shown in Figure 4, integrates an active MR-based optical AND gate (OAG) with supporting peripherals. Binary-encoded (fixed-point) operand values X_b and W_b are buffered, from which signs and magnitudes are extracted. The magnitudes are converted into stochastic bit-vectors X_V and W_V via B_to_S converters, serialized into bit-streams X and W at the target bitrate (BR), and driven into the OAG. Within the OAG, PN-doped terminals are modulated by the bit-streams to perform a bitwise AND, producing an optical pulse stream at the drop port ($DT(\lambda_{in})$). A complementary NAND pulse is generated at the through port ($TT(\lambda_{in})$). In Figure 4, switchable attenuators are integrated into the drop and through ports of the OAG. The modulators used from [21], [22] as the attenuators, deliver up to 10dB of optical power attenuation with nanowatt-scale dynamic power consumption and a compact 20μm device length [21]. The switching of the drop-port attenuator (at the top) is controlled by the XOR of $sign(X_b)$ and $sign(W_b)$, i.e., $sign(X_b{*}W_b)$. The switching of the through-port attenuator (at the bottom) is controlled by the inverse of $sign(X_b{*}W_b)$. The optical pulse streams at the drop and through ports are sent to compute-capable transducer units (PCAs) and then ADCs.

OSSM Operation. Figure 5(a) illustrates the passband shifts of the OAG MR under different operand inputs and temperature conditions. The MR's temperature, controlled via an integrated microheater and feedback circuit [23] (Figure 4), allows tuning its resonance from the fabrication-defined position η to a programmed position κ relative to the input wavelength λ_{in}. For each bit combination at the PN-doped operand terminals (X,W) = (0,1), (1,0), or (1,1), the MR's passband electro-refractively shifts to operand-driven positions (shown by red and blue passbands in Figure 5(a)). Based on these shifts relative to κ and λ_{in}, the MR's drop-port and through-port transmissions, $DT(\lambda_{in})$ and $TT(\lambda_{in})$, realize logical AND and NAND operations, respectively. Thus, the optical pulse streams at the drop and through ports represent bitwise AND and NAND results of the input bitstreams X and W.

To validate the operation of the OAG, we conducted a transient analysis, as shown in Figure 5(b)-(c). The OAG was modeled and simulated using foundry-validated tools from Ansys Lumerical's toolsuites [24]. Arbitrary bit-streams X and W (Figure 5(b)) were provided as inputs to the OAG model. The optical pulse streams generated at the drop and through ports were then measured (Figure 5(c)). As shown in the figure, the pulse stream at the drop port ($DT(AND)$) exhibits bitwise AND functionality, while the pulse stream at the through port ($TT(NAND)$) follows bitwise NAND functionality. When $sign(X_b{*}W_b)$ is '0' (i.e., positive multiplication result), the through port pulse stream is quenched by the bottom attenuator (see Figure 4), making the drop port pulse stream the positive stochastic multiplication result. On the other hand, when $sign(X_b{*}W_b)$ is '1' (i.e., negative multiplication result), the drop port pulse stream is quenched by the top attenuator (Figure 4). However, since the pulse stream at the through port represents bitwise NAND functionality, it does not directly correspond to stochastic multiplication. To address this, the pulse stream is inverted by the corresponding PCA' (Figure 4, explained further in the next Section 3.3) to generate the AND pulse stream (Figure 5(c)), which represents the stochastic multiplication result at the through port. As a result, OAGs can generate signed stochastic multiplication outcomes in the form of optical pulse streams.

3.3 Compute-Capable Transducer Units: PCAs

The stochastic multiplication bitstreams from the OAG are sent to compute-capable transducer units, namely PCA and PCA', as shown in Figure 4. Each consists of two stages: *(i)* optical-to-electrical transduction and *(ii)* analog pulse counting via a time-integrating receiver (TIR), followed by ADC digitization. The transduction stage uses a PD in PCA and a BPD in PCA'. These PD and BPD stages can undertake optical-to-electrical conversion with or without summing the incoming optical pulses. In both cases, a PD or BPD stage generates a train of electrical photocurrent pulses. PCA produces $f(DT(\lambda_{in}))$, while PCA' generates $f(logic'1') - f(TT(\lambda_{in}))$, an inverted NAND pulse stream enabling signed multiplication. The TIR integrates these pulses into an analog voltage proportional to their sum [27]. When a PD (or BPD) is not compute-capable, it generates a photocurrent pulse for each incident optical logic '1' pulse which accumulates a statistically significant analog voltage at the TIR output. Alternatively, a PD (or BPD) can inherently sum optical pulses via incoherent superposition when its sampling bandwidth exceeds the pulse rate [25], enabling incoherent superposition of optical pulses. If multiple pulses arrive within a window shorter than the inverse bandwidth, the resulting photocurrent represents the sum of β pulses—applicable to both homodyne and heterodyne signals [12], [25], [26].

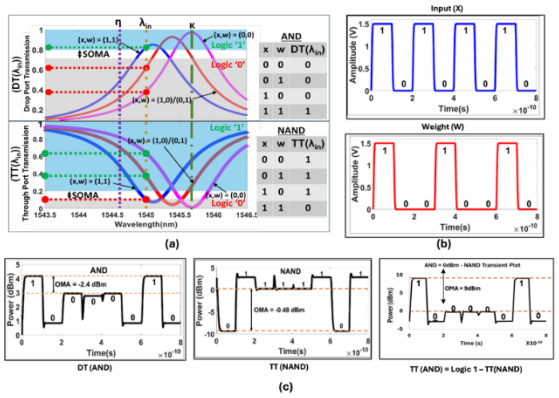


Figure 5: (a) Operation of OAG, (b) input X and weight W bit streams used for analysis, (c) results of OAG's transient analysis.

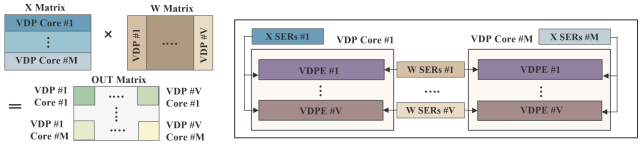


Figure 6: Dataflow used where the X and W matrices are mapped onto the VDPEs and VDP cores.

With b-bit precision, a signed stochastic multiplication generates 2^{b-1} optical pulses. Thus, if $\beta = 2^{b-1}$, each photocurrent pulse at a sampling event represents a single analog multiplication result. When $\beta > 2^{b-1}$, however, the photocurrent pulse reflects the analog sum of $floor(\beta \div 2^{b-1})$ multiplication results—effectively producing a VDP result between two vectors, each of size $floor(\beta \div 2^{b-1})$. Moreover, the subsequent TIR-based pulse counting stage can integrate up to $floor(\alpha \div 2^{b-1})$ photocurrent pulses, where $\alpha \gg \beta$, into a single analog voltage output. Operating the PD/BPD and TIR at the thermal noise floor can allow α to reach values as high as 10^7 [12], [25]-[27]. Therefore, by setting the sampling rate such that $\beta = 2^{b-1} = 128$, our OSSM can perform a temporal dot product of up to $\alpha \div 2^{b-1} = 78{,}125$ streaming X and W values.

3.4 Dataflow and Architectural Optimizations

Most prior photonic accelerators map one operand onto fixed photonic circuits that cannot be readily reconfigured, restricting these designs to a WS dataflow only [7]-[11]. In contrast, our VDP cores enable flexible dataflow selection by dynamically encoding both operands. We employ a fine-grained tiling strategy and meticulously designed spatial and temporal mappings. Our design utilizes an OS dataflow approach for performing GEMMs where OSSMs allow dynamic and fast switching of X and W operands. Furthermore, our design reduces on-chip buffer requirements and power consumption by enabling several B_to_S and serializer circuits to be shared across different VDPEs and VDP cores. As shown in Figure 6, matrix X is partitioned horizontally into M tiles, with each tile assigned to a separate VDP core. Each VDP core iteratively computes the results for one tiled row of X against the entirety of W. Similarly, matrix W is split into V columns, with each column mapped to the same VDPE across all VDP cores. Consequently, the B_to_S and serializer circuits for X are shared among all the VDPEs within a single VDP core, while those for W are shared across the same VDPEs among different VDP cores. Additionally, as discussed in Section 3.3, the PCAs perform temporal summation of numerous multiplications, allowing the ADCs to operate at MHz-range sampling rates. This reduction significantly lowers ADC area and power consumption.

4 Evaluation

4.1 Simulation Setup

We conducted comprehensive device- and architecture-level simulations to evaluate the proposed architecture's efficiency. Five transformer models (Table 1) were analyzed using a custom Python-based simulator that estimates performance and energy costs. The simulator performs layer-wise hardware-software mapping for each model and dataset, accurately modeling all peripherals and components (Table 2). Photonic signal losses and power consumption were assessed considering waveguide loss (1dB/cm [9]), splitter loss (0.13dB [28]), combiner loss (0.9dB [28]), MR through loss (0.02dB [11]), and OAG tuning/control power (6mW) with 3.5dB insertion loss [29]. A comb laser from [30] with >-3 dBm output across 25 usable wavelengths and 0.5W wall-plug power was used. Energy and performance estimates for ASTRA's LUTs and electronic buffers were derived using CACTI [31], while softmax and B_to_S converter circuits were synthesized using Xilinx Vivado. Model training and accuracy evaluations were conducted using PyTorch 2.3.

Our analysis shows that using 8-bit quantization for models results in inference accuracy comparable to that achieved with full precision (FP32), as shown in Table 3. Thus, we have selected transformer models with 8-bit precision, where the stochastic parameters are represented with 128 bits plus one sign bit.

4.2 Scalability and Error Analysis

We validated the scalability of our VDPE architecture through simulations using Lumerical Interconnect and Cadence Virtuoso. Results showed that each wavelength can support 1024 OAGs operating at 30Gbps with just ~0.5μW per OAG, enabling massive parallelism at low power. Moreover, we conducted an exhaustive design space exploration targeting minimal energy-delay product. The optimal configuration was identified as {M,V,N}={106,25,515}, where M is the number of VDP cores, V is the number of VDPEs per core, and N is the number of OAGs per VDPE. We also conducted an error analysis for our OSSM. It achieved a mean absolute error of 0.042, outperforming the stochastic multipliers reported in several previous efforts [17]. When applied to transformer model inference, the accuracy degradation, shown in Table 3 (Q(8-bit) + SC), is minimal.

4.3 Comparison with State-of-the-art

We compared ASTRA with CPU, GPU, TPU, and several transformer accelerators: an FPGA-based transformer accelerator FPGA_ACC [5], a processing-in-memory accelerator, TransPIM [6], MZM-based

Table 1: Transformer Model Configurations

Model	Params	Layers	N	Heads	d_{model}	d_{ff}
Transformer-base	52M	2	128	8	512	2048
BERT-base	108M	12	128	12	768	3072
Albert-base	12M	12	128	12	768	3072
ViT-base	86M	12	256	12	768	3072
OPT-350	350M	12	2048	12	768	3072

Table 2: Peripheral Parameters for ASTRA

Component	Latency	Power (mW)
Softmax	1.89ns	4.2664
LUT	0.2225ns	4.21
B_to_S	0.5302ns	0.021
Serializer [32]	0.03ns	1.5
ADC [33]	0.78ns	2.55
PCA	0.033ns~2.19ns	0.02
OSSM Attenuators [21]	~10 ps	10 nW

Table 3: Transformer Model Metrics

Model (metric)	Dataset	FP32	Q(8-bit)	Q(8-bit) + SC
Transformer-base	Ted-hrlr	70.90%	70.40%	70.10%
BERT-base	GLUE	87.00%	86.27%	85.98%
Albert-base	GLUE	86.07%	84.80%	84.51%
ViT-base	ImageNet	97.60%	96.50%	96.37%
OPT-350	Openassistant-Guanaco	18.07 (BLEU)	17.79 (BLEU)	17.49 (BLEU)

optical accelerator Lightning-Transformer (LT) [10], and an MR-based optical accelerator TRON [11].

4.3.1 Speedup Comparison. Figure 7 presents a speedup comparison between ASTRA, the various compute platforms, and the transformer accelerators evaluated. The speedup values are normalized against the CPU inference latency. On average, ASTRA achieves a speedup of 57314×, 6757×, 9567×, 1091×, 195×, 4.7×, and 7.6× over CPU, GPU, TPU, FPGA_ACC, TransPIM, TRON, and LT, respectively. The significantly lower latencies observed with ASTRA can be attributed to its high parallelism enabled by the proposed VDP core design, along with the integration of various device-, circuit-, and architectural-level optimizations, and the use of stochastic computing.

4.3.2 Energy Efficiency Comparison. The energy comparison results for ASTRA against the compute platforms and transformer accelerators are presented in Figure 8, with all energy values normalized to the CPU. ASTRA demonstrates average energy reductions of 1749.1×, 845.2×, 1254.1×, 9.1×, 3.9×, 2.6×, and 1.3× compared to the CPU, GPU, TPU, FPGA_ACC, TransPIM, TRON, and LT, respectively. These significant energy savings can be attributed

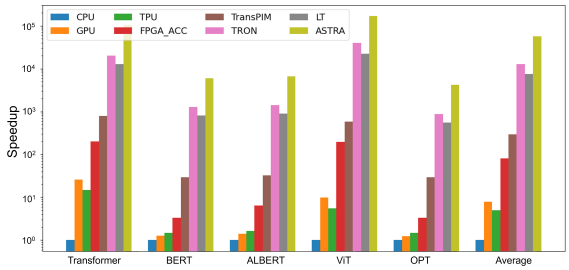


Figure 7: Speedup comparison.

to ASTRA's exceptionally low-latency operation in the optical domain, the reduced power consumption of its VDPEs enabled by the proposed low-cost OSSM design, and the elimination of DACs.

5 Conclusion

In this paper, we introduced a novel optical accelerator designed for LLMs. Our architecture integrates stochastic and analog computing while advancing the capabilities of optical VDP cores. The proposed optical homodyne VDPEs, incorporating stochastic signed multipliers, achieved significantly reduced latency and energy. Compared

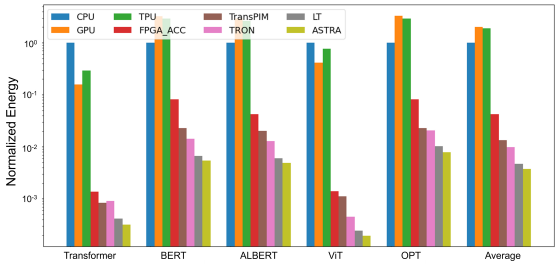


Figure 8: Energy comparison.

to GPU, TPU, CPU, and several state-of-the-art LLMs accelerators, we demonstrated at least 7.6× speedup and 1.3× lower energy consumption. These results highlight the potential of optical VDP cores combined with stochastic and analog computing for accelerating LLMs efficiently.

References

[1] A. Vaswani, *et al.* "Attention is all you need," NIPS, 2017.
[2] J. Devlin, *et al.* "BERT: pre-training of deep bidirectional transformers for language understanding," CoRR, Oct 2018.
[3] J. Achiam, *et al.*, "Gpt-4 technical report," arXiv preprint arXiv:2303.08774., 2023.
[4] A. Dosovitskiy, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, Oct. 2020.
[5] S. Lu, M. Wang, S. Liang, J. Lin, J. and Wang, Z. , "Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer," IEEE SOCC, 2020.
[6] M. Zhou, W. Xu, J. Kang and T. Rosing, "TransPIM: A Memory-based Acceleration via Software-Hardware Co-Design for Transformer," IEEE HPCA, 2022.
[7] F. Sunny, *et al.* "CrossLight: A cross-layer optimized silicon photonic neural network accelerator." ACM/IEEE DAC,2021.
[8] F. Sunny *et al.*, "RecLight: A Recurrent Neural Network Accelerator with Integrated Silicon Photonics." IEEE ISVLSI, 2022
[9] S. Afifi *et al.*, "GHOST: A Graph Neural Network Accelerator using Silicon Photonics." ACM, 2023
[10] H. Zhu, *et al.* "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator." IEEE HPCA, 2024.
[11] S. Afifi *et al.*, "Tron: Transformer neural network acceleration with non-coherent silicon photonics." GLSVLSI, 2023.
[12] S. Afifi, *et al.*, "Accelerating Neural Networks for Large Language Models and Graph Processing with Silicon Photonics", IEEE/ACM DATE, 2024.
[13] S. Li, "Scope: A stochastic computing engine for dram-based in-situ accelerator," IEEE/ACM MICRO, 2018.
[14] S. Afifi, *et al.*, "ARTEMIS: A Mixed Analog-Stochastic In-DRAM Accelerator for Transformer Neural Networks," IEEE/ACM CASES (ESWEEK), Oct 2024.
[15] S. S. Vatsavai, *et al.*, "SCONNA: A Stochastic Computing Based Optical Accelerator for Ultra-Fast, Energy-Efficient Inference of Integer-Quantized CNNs," IEEE IPDPS, 2023.
[16] S. S. Vatsavai and I. Thakkar, "A Bit-Parallel Deterministic Stochastic Multiplier," (ISQED), 2023.
[17] S. Mysore, *et al.*, "Atria: A bit-parallel stochastic arithmetic based accelerator for in-dram cnn processing," IEEE ISVLSI, 2021.
[18] Lightelligence, [Online]: https://www.lightelligence.ai/, Accessed on: Nov 15, 2024.
[19] Lightmatter, [Online]: https://lightmatter.co/, Accessed on: Nov 15, 2024.
[20] Luminous, [Online]: https://www.luminous.com, Accessed on: Nov 15, 2024.
[21] C. Ye, *et al.*, "λ-Size ITO and Graphene-Based Electro-Optic Modulators on SOI," IEEE Journal of Selected Topics in Quantum Electronics, 2014.
[22] J. K. George *et al.*, "Neuromorphic photonics with electro-absorption modulators," Opt. Express, OE, vol. 27, no. 4, pp. 5181–5191, Feb. 2019.
[23] T. Ferreira de Lima *et al.*, "Design automation of photonic resonator weights," Nanophotonics, vol. 11, no. 17, pp. 3805–3822, 2022.
[24] "Pic design and simulation software- lumerical interconnect," Apr 2021. [Online]. Available: https://www.lumerical.com/products/interconnect/
[25] F. Brückerhoff-Plückelmann *et al.*, "A large scale photonic matrix processor enabled by charge accumulation," Nanophotonics, vol. 12, no. 5, pp. 819–825, Mar. 2023.
[26] S. S. Vatsavai, *et al.*, "An Optical XNOR-Bitcount Based Accelerator for Efficient Inference of Binary Neural Networks," ISQED, 2023.
[27] A. Sludds, *et al.*, "Delocalized photonic deep learning on the inter net's edge," Science, 2022.
[28] L. H. Frandsen, *et al.*, "Ultralow-loss 3-dB photonic crystal waveguide splitter," Optics letters 29, 14, 1623-1625, 2004.
[29] V. S. Praneeth Karempudi, S. Sri Vatsavai, I. Thakkar, and J. T. Hastings, "A Polymorphic Electro-Optic Logic Gate for High-Speed Reconfigurable Computing Circuits," ISQED, Apr. 2023.
[30] A. Rizzo *et al.*, "Massively scalable Kerr comb-driven silicon photonic link," Nat. Photon., vol. 17, no. 9, pp. 781–790, Sep. 2023.
[31] HP Labs : CACTI. [Online]: https://www.hpl.hp.com/research/cacti/.
[32] S. Lin, *et al.*, "Electronic–Photonic Co-Optimization of High-Speed Silicon Photonic Transmitters," Journal of Lightwave Technology, 2017.
[33] D.-R. Oh *et al.*, "An 8b 1gs/s 2.55mw sar-flash adc with complementary dynamic amplifiers," IVLSIC, 2020.