Accepted: 07 September 2025
Received: 22 July 2025

RESEARCH-ARTICLE

ASTRA: A Stochastic Transformer Neural Network Accelerator with Silicon Photonics

SALMA AFIFI, Colorado State University, Fort Collins, CO, United States

OLUWASEUN ALO, University of Kentucky, Lexington, KY, United States

ISHAN G THAKKAR, University of Kentucky, Lexington, KY, United States

SUDEEP PASRICHA, Colorado State University, Fort Collins, CO, United States

Open Access Support provided by:

University of Kentucky

Colorado State University

ASTRA: A Stochastic Transformer Neural Network Accelerator with Silicon Photonics

SALMA AFIFI

Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado, United States, salma.afifi@colostate.edu

OLUWASEUN ALO

University of Kentucky, Lexington, Kentucky, United States, seun.alo@uky.edu

ISHAN THAKKAR

University of Kentucky, Lexington, Kentucky, United States, igthakkar@uky.edu

SUDEEP PASRICHA

Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado, United States, sudeep@colostate.edu

Transformers have emerged as a dominant architecture in deep learning, demonstrating unparalleled success across a wide range of applications, including natural language processing (NLP), computer vision (CV), and scientific computing. By leveraging the self-attention mechanism, transformers achieve superior performance over traditional models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs). However, these performance gains come at a cost—high computational complexity and substantial memory requirements, making transformers particularly challenging to deploy efficiently on conventional hardware. To address the increasingly intensive computational demands of attention-based transformers, there is growing interest in developing efficient and high-speed hardware accelerators. Silicon photonics has emerged as a promising alternative to digital electronics, offering high-bandwidth and low-latency computation while improving overall computational and energy efficiency. This work introduces ASTRA, the first optical hardware accelerator that leverages stochastic computing principles for transformer neural networks. ASTRA incorporates novel full-range optical stochastic multipliers and stochastic-analog compute-capable optical-to-electrical transducer units to efficiently handle both static and dynamic tensor computations in attention-based models. Through detailed performance analysis, we demonstrate that ASTRA achieves at least 7.6× speedup and 1.3× lower energy consumption compared to state-of-the-art transformer accelerators.

CCS CONCEPTS • Computer systems organization~Architectures~Other architectures~Optical computing • Hardware~Very large scale integration design~Application-specific VLSI designs • Computing methodologies~Machine learning~Machine learning approaches~Neural networks

Additional Keywords and Phrases: Transformer neural networks, silicon photonics, inference acceleration, stochastic computing, optical computing.

1 INTRODUCTION

Transformers have rapidly become foundational to advancements in natural language processing (NLP) and computer vision, driving breakthroughs in tasks such as machine translation, question-answering, and image recognition. By making use of powerful attention mechanisms, transformers excel at learning complex, long-distance dependencies within data, resulting in highly accurate and context-aware outputs [1]. Pioneering models such as BERT [2], GPT [3], and Vision Transformers (ViTs) [4] have set new performance benchmarks across diverse applications. However, this remarkable accuracy comes at a substantial computational cost: transformers often comprise billions of parameters, leading to surging inference latencies, energy consumption, and processing complexity. These constraints underscore an urgent need for specialized hardware accelerators that can efficiently manage the unique computational workloads of transformers.

ACM 1539-9087/2025/09-ART
http://dx.doi.org/10.1145/3769092

To address these challenges, several efforts have been focusing on the development of hardware accelerators designed to improve transformer inference efficiency. Several digital electronic accelerators have been proposed, incorporating transformer-specific architectural optimizations to improve throughput and reduce latency [5]-[8]. Despite these advances, traditional digital computing platforms face significant hurdles as transistor-based chips approach the limits of Dennard scaling, leading to increased power dissipation per unit area and diminishing performance gains [9]. This has led researchers to explore alternative technologies. One such promising technology is silicon photonics, which can enable ultra-fast light-speed data transmission, high levels of parallelism, and energy efficiency.

Silicon photonic accelerators have shown promise in accelerating neural network operations by leveraging high-bandwidth optical devices such as Mach-Zehnder modulators (MZMs) and microring resonators (MRs) [10]-[16]. These accelerators have been applied to various deep neural networks (DNNs), and several startups are integrating optical computing into their commercial products [33]-[36]. However, despite their advantages, existing photonic accelerators face critical challenges that hinder scalability and practical deployment. A key challenge is high insertion loss, as optical signals passing through multiple MR or MZM devices experience significant attenuation. Additionally, heterodyne crosstalk noise, a major issue in wavelength-division multiplexing (WDM) systems, limits the number of usable wavelengths per vector dot product engine (VDPE), constraining scalability. Many designs also rely on power-hungry digital-to-analog converters (DACs) to tune the MRs, increasing cost and complexity. Moreover, existing photonic accelerators predominantly enforce a weight-stationary (WS) dataflow. While WS is effective for convolutional neural networks (CNNs) and recurrent neural networks (RNNs), it is poorly suited for transformer neural network models, which require dynamic operand generation for attention computations. The inefficiency of WS in transformers exacerbates performance limitations, restricting the applicability of conventional optical architectures from prior work.

Recent works such as TRON [15] and Lightning-Transformer [14] have attempted to address these challenges with non-coherent optical accelerator designs optimized for transformers. TRON reduces opto-electronic conversions to improve inference efficiency but remains constrained by heterodyne crosstalk and the need to duplicate VDPEs to handle positive and negative values, increasing resource overhead. Lightning-Transformer employs a photonic tensor core based on interference-driven VDPEs using MZMs. However, this approach introduces several challenges, most notably sensitivity to phase noise which necessitate additional mechanisms to ensure correct and error-free operation [47].

To overcome these limitations of the state-of-the-art, we introduce ASTRA, the first optical accelerator to leverage stochastic computing for transformer neural network acceleration. ASTRA employs a novel optical stochastic signed multiplier (OSSM), fundamentally rethinking optical-domain transformer acceleration by replacing amplitude-based analog encoding with stochastic computation. This approach reduces optical dynamic range, and crucially, removes the need for DACs, significantly improving power efficiency. Furthermore, ASTRA partitions each VDP core into wavelength-specific VDPEs, allowing each VDPE to operate independently on a single wavelength. This eliminates heterodyne crosstalk. While stochastic computing has traditionally suffered from high error rates, ASTRA overcomes this accuracy challenges by employing a low-error, low-cost multiplication technique that efficiently performs bitwise operations in the optical domain. Additionally, ASTRA avoids stochastic additions, which typically introduce significant inaccuracies, by instead leveraging a temporal analog accumulation approach. This strategy not only enhances precision but also minimizes data movement overhead, enabling fast and accurate successive data accumulations. By integrating stochastic encoding with massively parallel OSSMs—each VDPE supporting up to 1024 OSSMs—ASTRA achieves a scalable, high-bandwidth optical accelerator tailored to the dynamic and dataflow-intensive computations of transformers.

In this paper, we present ASTRA's architecture and its ability to efficiently harness the advantages of optical computing while mitigating key challenges that have hindered prior designs. Our novel contributions are:

- We propose the first silicon-photonic based hardware accelerator that integrates stochastic and analog computing tailored for transformers to achieve high performance and energy efficiency.
- 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 substantially 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.
- We propose several device-, architecture- and circuit-level optimizations to lower input-to-optical and analog-to-digital conversion costs, and to support low-latency optical computations.
- We perform a comprehensive comparison with GPU, TPU, CPU, and several state-of-the-art hardware accelerators for transformers.

2 BACKGROUND

2.1 Transformer Neural Networks

Transformer neural networks, initially proposed for sequence transduction tasks in NLP, have become a foundational model in machine learning, particularly in tasks that require learning long-term dependencies [1]. The key innovation in transformers is the attention mechanism, which allows the model to compute pairwise correlations across the entire input sequence, enabling the handling of long-range dependencies more efficiently than traditional recurrent models. The core architecture of a transformer consists of encoder and decoder blocks. The encoder processes a given input sequence to generate a continuous, high-dimensional representation, while the decoder iteratively generates output tokens based on both the encoded representation and the previously produced outputs. The encoder and decoder blocks consist of two primary sub-blocks: multi-head self-attention (MHA) and feed-forward network (FFN) as shown in Figure 1.

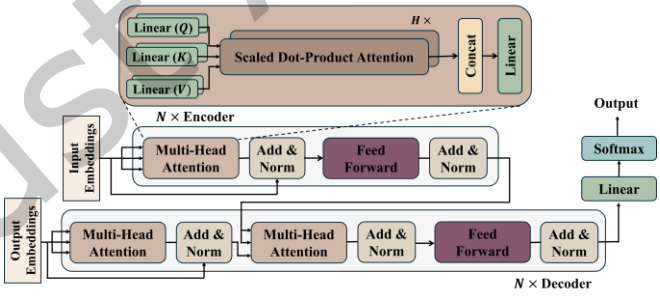


Figure 1: Transformer neural network architecture overview.

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

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

where d_k is the dimension of Q and K. The attention mechanism produces its output by concatenating the results from multiple attention heads and then applying a linear transformation.

The FFN typically consists of two dense layers with an activation function, such as GELU or ReLU, applied between them. More recent transformer-based pre-trained language models, such as BERT [2] and its variants [3], utilize a stack of transformer encoder blocks exclusively, arranged in N cascaded layers, followed by a feed-forward layer, GELU activation, and normalization layers. Similarly, the ViT [4] comprises N encoder layers followed by a multi-layer perceptron, where the input sequence vectors represent segments of an image. In contrast, models like GPT-4 [3] employ only decoder blocks in their architecture.

While transformers 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 transformer models a highly challenging endeavor.

2.2 Stochastic Computing

The stochastic computing (SC) paradigm reduces computational complexity by representing values with sequences of individual bits, trading precision for simpler logic design and lower static power consumption. Due to these efficiencies, SC has gained traction in areas such as image and signal processing, control systems, DNNs, and general-purpose computing [18], [19]. In stochastic computing, real numbers are represented through probabilistic bit-streams, where the occurrence rate of 1s to 0s reflects the real value. Eq. (2) and (3) below outline examples of stochastically representing two binary numbers:

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

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

After this encoding step, computations are performed by statistically manipulating the input bit-streams, allowing many standard arithmetic functions to be performed with simple logic gates rather than complex circuits [18]-[22]. 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 ASTRA accelerator introduces specialized methods to overcome such inaccuracies and enhance computational precision.

SC incurs a storage overhead of O(2^n), as representing an n-bit real value requires 2^n bits in a stochastic format. To address this, SC approaches typically store operands in a binary format, necessitating stochastic-to-binary (S_to_B) conversion before processing. This conversion is commonly performed using a population count (PC) unit, which counts the number of 1's in a stochastic bitstream to determine its corresponding binary value. However, PC units introduce significant challenges, including high area, latency, and energy consumption [18], [19]. ASTRA overcomes these limitations by employing a low-overhead S_to_B conversion technique, optimizing both computational efficiency and hardware complexity.

2.3 Optical Analog Computations for ANN Acceleration

Optical ANN accelerators have attracted significant interest from both academic and industry researchers due to their high performance and energy efficiency benefits [10]-[16], [33]-[36]. These accelerators are usually designed to operate in either a coherent or non-coherent manner. Coherent architectures encode

parameters onto the optical signal's phase to execute multiply and accumulate (MAC) operations [22]. In contrast, non-coherent architectures imprint parameters onto the optical signal's amplitude. Multi-wavelength optical signals enable parallel operations with banks of opto-electric modulators typically based on microring resonator (MR) devices operating in parallel. MRs are used to alter the optical signal's amplitude and enable optical computations. Each MR can be designed and tuned to work at a specific wavelength, referred to as the MR's resonant wavelength (λ_{MR}), defined as:

$$\lambda_{MR} = \frac{2\pi R}{m} n_{eff} \tag{5}$$

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 VDP core, and also shows an MR modulator in the activation bank (corresponding to λ_{MR1}) imprinting an activation value onto the signal transmission.

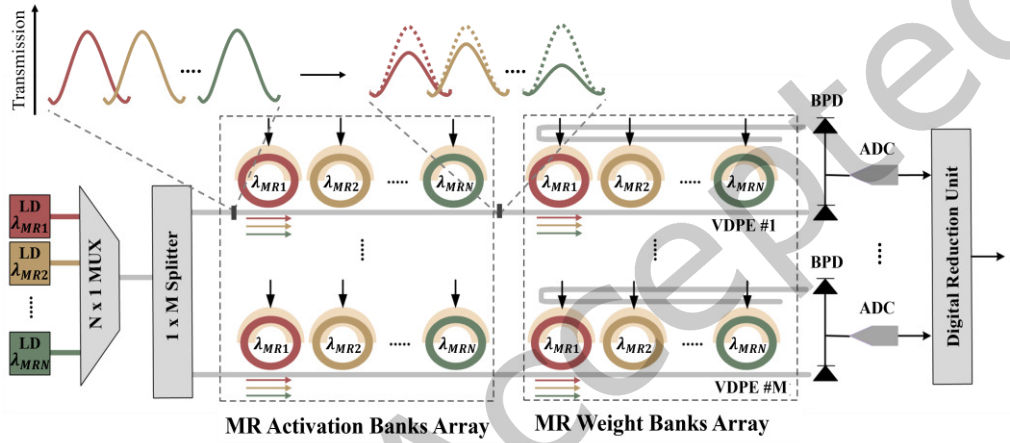


Figure 2: An optical VDP core used to perform the matrix multiplication between activation and weight matrices. The optical signal transmission is shown at the first MR's input and through ports before and after imprinting an activation.

Non-coherent optical accelerators typically use wavelength-division multiplexing (WDM) to boost throughput and simulate neuron functionality in artificial neural networks (ANNs). This approach combines multiple optical signals of distinct wavelengths within a single waveguide through an optical multiplexer [15]. The waveguide is designed to pass near an array of MRs, each tuned to a specific wavelength in the waveguide, enabling parallel multiplications.

Figure 2 shows the architecture of a VDP core used in most non-coherent optical accelerators [10]-[13]. Here, N wavelengths are used where each unique wavelength (shown with a different color) is emitted using a laser diode (LD). The wavelengths are multiplexed into a single waveguide and then divided into M branches using a $1 \times M$ splitter. VDP operations are computed across the vector dot-product elements (VDPEs), shown as separate lanes in the figure (VDPE #1 ... VDPE #M), where each VDPE performs N multiplications concurrently. The VDPE structure is divided into two MR bank arrays. The first MR bank encodes N activation values onto N distinct wavelengths, with each MR tuned to operate at a specific wavelength. The activation-encoded optical signals then proceed through the second MR bank, where corresponding weight values are similarly encoded on these signals, resulting in N parallel multiplication operations. A balanced photodetector (BPD) then sums the outputs across positive and negative weight arms for each branch in the weight bank. The resulting analog output is converted to a digital signal using and analog-to-digital converter (ADC), which is sent to the digital reduction unit for summing all partial sums from each VDPE.

2.4 Prior Work and Motivation

While the architecture depicted in Figure 2 offers significant parallelism as quantified in [10]-[13], several limitations impede its scalability and the full utilization of the high bandwidth available in the optical domain. First, the insertion losses in optical signals can be very high as every optical signal per VDPE faces the insertion loss of 2 in-band MRs and $2 \times (N - 1)$ out-of-band MRs. Second, heterodyne crosstalk is a significant challenge and one of the primary sources of noise in WDM systems. It occurs when the signal power at one wavelength is impacted by noise power from one or more different wavelengths in the utilized spectrum. Detailed device-level analysis in [13] suggests that to maintain accuracy at 8-bit precision, a maximum of 18 wavelengths (or 36 MRs per VDPE) can be used. Other studies [31] [32] suggest that if shot noise and thermal noise are also dominant in a VDPE in addition to the heterodyne crosstalk noise at 8-bit precision, the number of supported wavelengths can be as low as 1 per VDPE (2 MRs per VDPE). This constraint limits scalability, necessitating additional VDP cores, and digital circuitry for reductions. Third, a significant limitation is the heavy reliance on digital-to-analog (DAC) devices which are required to convert the model parameters to analog signals to tune the MRs, resulting in high area/power cost and complexity. Lastly, in state-of-the-art optical neural network accelerator designs, weights are typically mapped statically to the MR banks, restricting the dataflow to weight stationary (WS) and limiting efficient support for output stationary (OS) or input stationary (IS) dataflows. While these constraints have been manageable for optical accelerators tailored to ANNs such as CNNs, RNNs, and GNNs, they become more pronounced for transformer neural networks. For instance, the dynamic and complex nature of attention computations, where operands for matrix multiplications are generated at runtime, often renders WS dataflows inefficient for many scenarios.

A recent work introduced a non-coherent optical accelerator called TRON [15], designed to accelerate transformer neural networks through architecture-level optimizations aimed at reducing opto-electronic conversions during inference. However, its throughput remains constrained by challenges such as heterodyne crosstalk and the necessity to duplicate VDPEs for handling both positive and negative values, leading to increased resource overhead and inefficiencies. Lightning-Transformer [14] is another non-coherent optical accelerator that introduces a dynamically-operated photonic tensor core (DPTC), which consists of interference-based crossbar arrays of optical VDPEs that utilize MZMs. Unlike traditional weight-static photonic accelerators, Lightning-Transformer enables dynamic matrix multiplications by encoding both operands in the optical domain, eliminating the need for pre-programmed weights. This design allows for high-speed operand switching and minimizes latency caused by slow device reconfigurations. However, despite its optimizations, Lightning-Transformer still requires additional MZMs to encode sign bits on the signal phase before multiplication, increasing system complexity. Moreover, DAC power consumption remains a dominant bottleneck, particularly in higher-precision computations, limiting overall efficiency. Other electronic-based transformer accelerators have primarily focused on optimizing specific transformer models or targeting individual layers for acceleration. For example, the work in [8] introduced an FPGA-based accelerator designed to enhance the efficiency of MHA and FF layers by partitioning weight matrices in a way that enables shared hardware resources between the two layers. Similarly, [6] proposed another FPGA-based framework that incorporates structured pruning and specialized storage techniques to handle sparse matrices efficiently.

Beyond FPGA-based approaches, TransPIM [7] leverages processing-in-memory (PIM) and near-memory computing (NMC) to accelerate transformers, addressing the inefficiencies of traditional architectures that suffer from excessive data movement and limited parallelism. Unlike conventional memory-based accelerators that rely on layer-based dataflows, TransPIM introduces a token-based dataflow, which significantly improves data locality by keeping computations of related tokens within the same memory location. At the hardware level, TransPIM integrates lightweight modifications into DRAM-based high-bandwidth memory (HBM), introducing auxiliary computing units (ACUs) that enable efficient vector

reductions and softmax calculations—two key bottlenecks in transformer models. Similarly, HAIMA [45] demonstrates the benefits of token-based dataflows through a hybrid SRAM–DRAM architecture, optimizing both matrix multiplications and data transfers. However, both DRAM- and SRAM-based PIM architectures require digital implementations of MAC operations. These are typically decomposed into multiple memory operation cycles (MOCs) [19], thereby increasing the computational overhead for MAC-intensive workloads in state-of-the-art in-DRAM accelerators. Moreover, digital in-memory realization of complex functions like reduction and softmax within bit-cell arrays remains a non-trivial challenge. In addition to these architectural complexities, effectively managing dataflows, scheduling, and coordinating operations in both PIM and NMC environments introduces further system-level design challenges. Although PIM designs based on non-volatile memories (NVMs)—such as ReRAM in [46]—can mitigate some of these limitations, they bring their own set of concerns. NVM-based architectures often suffer from limited endurance, reliability issues, and security vulnerabilities. Frequent in-memory computation may accelerate wear-out, reducing device lifespan. Furthermore, process variations, thermal fluctuations, and error accumulation can impair computational accuracy and overall system reliability [19].

SCONNA [20] is a prior effort that explores optical stochastic computing for CNNs. While both SCONNA and ASTRA employ stochastic principles in the optical domain, their target workloads and architectural designs diverge significantly. SCONNA was developed for CNNs with static weight reuse and regular operand patterns, enabling unsigned stochastic multiplication and statically mapped dataflows. In contrast, transformer models require dynamic generation of Q/K/V matrices and runtime-varying matrix multiplications. These demands necessitate fundamentally different capabilities, including signed computation, dynamic operand scheduling, and scalable interconnects, capabilities that ASTRA introduces through a novel design tailored specifically for transformers. Furthermore, ASTRA introduces a novel homodyne, single-wavelength VDPE design that eliminates heterodyne crosstalk and minimizes insertion losses, a departure from SCONNA's traditional WDM-based cascaded MR approach, which suffers from scalability and noise limitations

In contrast to these existing solutions, our proposed ASTRA architecture in this work introduces a novel optical computing substrate based on optical stochastic signed multipliers (OSSMs). Each VDP core in ASTRA is partitioned into wavelength-specific VDPEs, with each VDPE operating on a single wavelength, effectively eliminating heterodyne crosstalk. ASTRA leverages the OSSMs to encode single-bit values (two power levels – ON and OFF) onto optical signals in a stochastic format. The use of ON-OFF power levels significantly reduces the optical dynamic range compared to analog multi-level optical signals used in non-coherent analog photonic accelerators, lowering optical power requirements for ASTRA. Moreover, the use of a stochastic number format eliminates the need for DACs. Conventional optical multiplication is typically performed using quantized multi-bit representations, requiring complex optical encoding and computation. In contrast, ASTRA simplifies this process by reducing multiplication to a bitwise AND operation between ON-OFF power level encoded stochastic bitstreams using the OSSMs. This approach streamlines computation, minimizes optical complexity, and enhances parallelism. As discussed in Section 4.2, each VDPE can accommodate up to 1024 OSSMs, enabling massively parallel processing optimized for the dynamic computational demands of transformers.

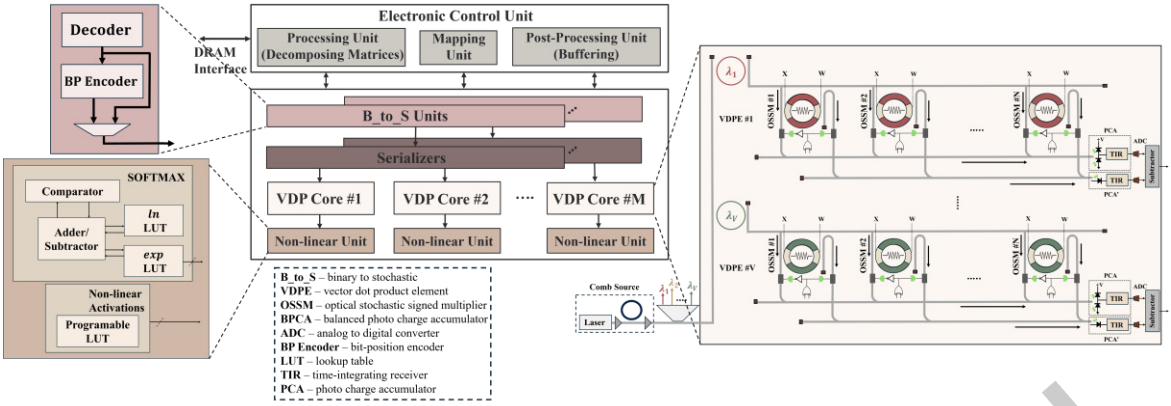


Figure 3: ASTRA architecture overview showing the vector dot-product (VDP) cores, non-linear units, binary-to-stochastic (B-to-S) circuits, and serializers.

3 ASTRA HARDWARE ACCELERATOR

In this section, we present an overview of ASTRA, our optical transformer accelerator, with an architectural overview illustrated in Figure 3. The main computational tasks, specifically the general matrix multiplications (GEMM), are executed within the VDP cores. Stochastic operands are generated and delivered to the VDP cores through binary-to-stochastic (B_to_S) converter units and high-speed serializers. An integrated electronic-control unit (ECU) manages the interfacing with main memory, buffering of operands, resource allocation, and mapping of weight matrices to the photonic architecture. The following subsections describe the ASTRA architecture along with the device-, circuit-, and architecture-level optimizations implemented to efficiently accelerate transformer neural networks.

3.1 VDP Cores

The VDP cores in ASTRA employ a novel architectural design, departing from the conventional non-coherent wavelength division multiplexing (WDM) based approach discussed in Section 2.3. A low-power laser comb source generates V wavelengths that are coupled into a waveguide. These wavelengths are 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 [18]-[22], ASTRA employs a low-cost and low-error deterministic B_to_S conversion method [23]. 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, which is part of the B_to_S unit shown 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 resulting at the output exhibit bit-position correlation so that the conditional probability of finding 1s in the first bit-vector given the second bit-vector is approximately equal to the marginal probability of finding 1s in the first bit-vector. Meeting this probability condition ensures minimal errors and accuracy drops for stochastic multiplications, as demonstrated in [21] and [23]. The two stochastic bit-vectors are then serialized and multiplied using the OSSM units that reside in VDP cores.

The homodyne (i.e., carried by the same wavelength) optical output streams of all OSSMs in a VDPE are aggregated through separate lanes: one for positive values and another for negative values. Photodetectors (PDs) employ homodyne incoherent superposition and charge accumulation [28] to convert these aggregated optical streams (multiple, parallel streams) into a stream of electrical analog pulses. This optical-to-electrical conversion process also inherently implements two arithmetic functions in parallel: *(i)* population count (PC) of stochastic streams, and *(ii)* accumulation of stochastic multiplication results. Thus, the output stream of electrical analog pulses from the PDs is the combined result of the optical-to-electrical conversion and the

above two arithmetic functions. These electrical pulses are then added in the analog format to the previously accumulated pulses in the photo-charge accumulators (PCA) for positive values and PCA' for negative values. Once all the analog pulses are accumulated and partial sums have been generated, ADCs convert the accumulated values from the PCA and PCA' into digital form, and a digital subtractor is employed to compute the final result.

The VDP core design achieves high efficiency by utilizing optical computing while overcoming key limitations of traditional architectures, such as costly DAC arrays and S_to_B converters [18]-[22]. In traditional stochastic platforms, popcount (PC) units convert stochastic bitstreams to binary by counting '1's but they are limited by high area, latency, and energy costs [18], [19]. These PC units do not support in-situ accumulation of operands, or the simultaneous S_to_B conversion of multiple parallel bistreams. In contrast, ASTRA combines the S_to_B conversions of multiple parallel bitstreams with their accumulation and optical-to-electrical conversion more efficiently through the PCA structures and ADCs. Our architectural design is also able to eliminate heterodyne crosstalk as each VDPE employs a single wavelength channel, and coherent crosstalk is not present due to the incoherent operation of VDPEs. Furthermore, optical insertion losses are significantly reduced since each optical signal passes through only 2 in-band MRs: one filter MR at the beginning of the VDPE and one OSSM. This is a substantial improvement compared to conventional optical VDPE designs, where an optical signal typically traverses 2 in-band MRs and 2×(N-1) out-of-band MRs, as discussed in Section 2.3. The following subsections describe the novel OSSM and PCA structures within each VDPE.

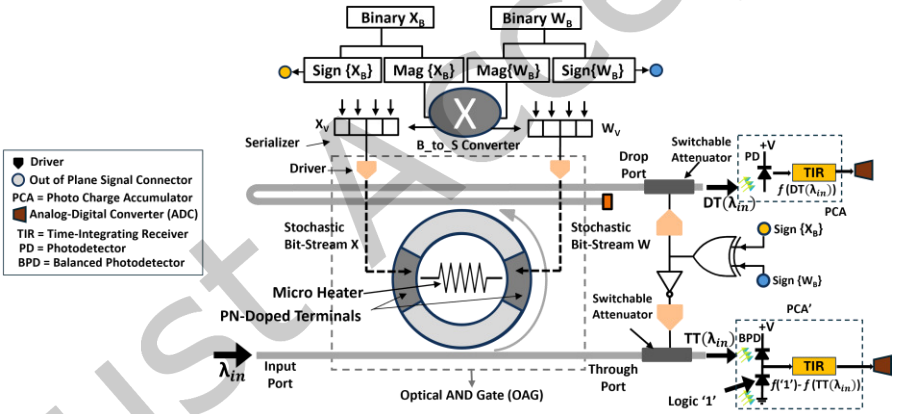


Figure 4: Schematic of our Optical Stochastic Signed Multiplier (OSSM).

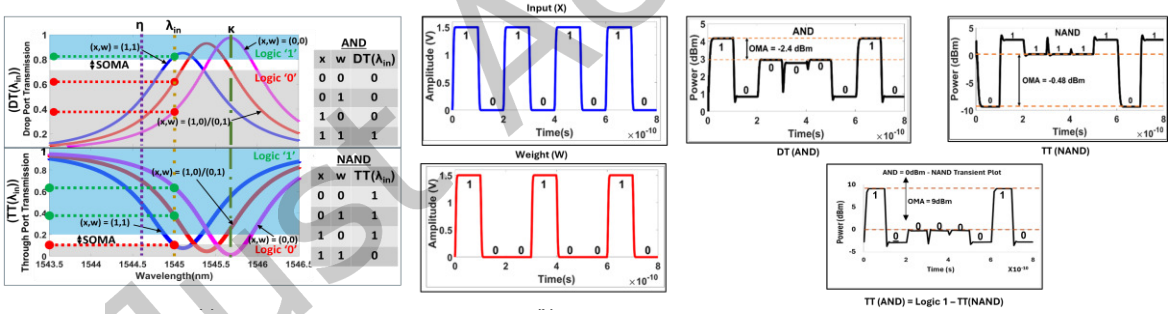
3.2 Optical Stochastic Signed Multiplier (OSSM): Structure

Our Optical Stochastic Signed Multiplier (OSSM), shown in Figure 4, employs an active MR-based optical AND gate (OAG) and peripherals. The peripherals include a buffer that provides the binary-encoded (fixed-point precision) values of operands X_b and W_b. The signs and magnitudes of these fixed-point binary-encoded values are first extracted. The magnitudes of the input values are then converted into stochastic bit-vectors X_V and W_v by the B_to_S converters. These bit-vectors are serialized into stochastic bit-streams X and W at the target bitrate (BR) using high-speed serializers, and subsequently provided as input to the OAG via high-speed drivers. In the OAG, the PN-doped terminals are driven by the incoming bit-streams, enabling a bitwise AND operation that produces the optical pulse-stream at the drop port, corresponding to the drop-port transmission ($DT(\lambda_{in})$ in Figure 4). The byproduct pulse stream, generated as the through port transmission ($TT(\lambda_{in})$), represents bit-wise NAND results. The operation of the OAG that enables it to generate these optical bit-streams is explained in Section 3.3.

In Figure 4, switchable attenuators are integrated into the drop and through ports of the OAG. To ensure energy efficiency, we utilize electro-absorption modulators from [25], [26] as the attenuators. These modulators deliver up to 10dB of optical power attenuation with nanowatt-scale dynamic power consumption and a compact 20μm device length [25]. The switching of the drop-port switchable 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 switchable attenuator (at the bottom) is controlled by the inverse of $sign(X_b * W_b)$, using the inverter shown in Figure 4. The optical pulse streams at the drop and through ports are sent to compute-capable transducer units (PCAs) and then ADCs.

3.3 Optical Stochastic Signed Multiplier (OSSM): Operation

Figure 5(a) illustrates the passband positions of the OAG MR for various operand inputs and temperature conditions. The MR temperature can be raised using the integrated microheater (Figure 4), allowing the resonance to shift from its initial fabrication-defined position η to the programmed position κ, relative to the input optical wavelength λ_{in}. Figure 5(a) shows the passband positions of the OAG MR for different operand inputs and temperature conditions. The temperature of the MR can be increased using the integrated microheater and a feedback control circuit [37], tuning the operand-independent resonance from its fabrication-defined initial position η to the programmed position κ relative to the input optical wavelength position λ_{in} (Figure 5(a)). For each bit combination at the PN-doped operand terminals (X, W) = (0,1), (1,0), or (1,1), the resonance passband of the MR electro-refractively shifts to an operand-driven position (shown by the red and blue passbands in Figure 5(a)). Based on the spectral positions of the shifted MR passbands, relative to the spectral positions κ and λ_{in}, the drop-port transmission $DT(\lambda_{in})$ and through-port transmission $TT(\lambda_{in})$ of the MR follow the truth tables of logical AND and logical NAND operations, respectively. Thus, the optical pulse stream at the drop port (through port) provides bitwise logical AND (logical NAND) results between the input stochastic bit-streams X and W.



(a) (b) (c)

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

To validate the operation of the OAG, we conducted a transient analysis, the results of which are shown in Figure 5(b)-(c). The OAG was modeled and simulated using foundry-validated tools from Ansys Lumerical's Device, Charge, and Interconnect tool suites [27]. 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, Section 3.4) 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.

3.4 Compute-Capable Transducer Units: PCAs

The stochastic multiplication bit-streams generated by the OAG are directed to compute-capable transducer units. Two different designs of transducer units are used in Figure 4, namely PCA and PCA'. Both PCA and PCA' have two compute-capable stages: *(i)* an optical-to-electrical transduction stage, and *(ii)* an analog pulse counter stage that consists of a time-integrating receiver (TIR). After these stages, an ADC is used. The optical-to-electrical transduction stage employs 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. In Figure 4, $f(DT(\lambda_{in}))$ and $f(logic'1')$-$f(TT(\lambda_{in}))$ are the trains of electrical photocurrent pulses generated by the PD of PCA and BPD of PCA', respectively. The photocurrent pulse train $f(logic'1')$-$f(TT(\lambda_{in}))$ corresponds to the inverted NAND pulse stream discussed in the previous section, which enables signed multiplication at the through port. The TIR-based pulse counter stage integrates the incoming photocurrent pulses to generate an analog voltage output, proportional to the sum of the incident photocurrent pulses [30]. Thus, the PD/BPD and TIR-based pulse counter stages perform summation.

When a PD (BPD) is not compute-capable, it generates a photocurrent pulse for each optical logic '1' pulse incident upon it. The amplitude of a photocurrent pulse generated for an optical logic '0' remains under the noise limit; therefore, a logic '0' remains statistically undetected. The current pulse generated by an optical logic '1' accrues a certain statistically significant amount of analog voltage at the output of the TIR. Alternatively, a PD (BPD) unit can be made to inherently compute the summation of incoming optical pulses by setting its sampling bandwidth to be higher than the arrival rates of the optical pulses [28]. In this case, an incoherent superposition occurs for all the incoming optical pulses that arrive in a period that is shorter than the inverse sampling bandwidth of the PD (or BPD) [28]. Such incoherent superposition can occur between homodyne and heterodyne optical pulses [20], [28], [29]. The resultant photocurrent pulse at each sampling event, therefore, represents a sum of multiple (suppose β) optical pulses if all β pulses arrive in a period that is shorter than the inverse bandwidth of the PD (or BPD).

Since a signed stochastic multiplication result of b-bit precision will have 2^{b-1} optical pulses, the generated photocurrent pulse at each sampling event would represent an analog-converted multiplication result when $\beta = 2^{b-1}$. In another case, if $\beta > 2^{b-1}$, the generated photocurrent pulse at each sampling event would represent the analog sum of a total of $floor(\beta \div 2^{b-1})$ multiplication results; thus, it would represent a VDP result between the vectors of size $floor(\beta \div 2^{b-1})$ each. The subsequent TIR-based pulse count stage can enable the summation of a massive number of optical pulses $\alpha \gg \beta$, by allowing the integration of a total of $floor(\alpha \div 2^{b-1})$ photocurrent pulses into the resultant analog voltage output. If the PD/BPD and TIR are operated at the thermal noise floor, the feasible value of α can be as high as 10^7 [20], [28]-[30]. As a result, if we set the sampling rate of our PD/BPD units to achieve $\beta = 2^{b-1} = 128$, our OSSM can perform a temporal dot-product (temporal sum of multiplications) of a total of $\alpha \div 2^{b-1} = 78{,}125$ optically streaming X and W values.

3.5 Non-linear Activation Units

While some previous optical accelerators implement non-linear functions directly in the optical domain, they often require digital-to-analog converters (DACs) and vertical-cavity surface-emitting lasers (VCSELs) to convert outputs back into optical signals, incurring significant power and latency overhead. In contrast,

ASTRA's vector dot product engines (VDPEs) output results in the digital domain, allowing us to implement non-linear functions entirely using simple lookup tables (LUTs) and digital circuits. This eliminates the need for additional costly optical-electric data conversions, reducing both power consumption and execution time.

Non-linear activation functions (Figure 3) such as ReLU and GELU, which are usually required in FFN blocks, can be directly realized with dedicated LUTs. However, the softmax function, frequently used in MHA layers, presents additional challenges. Specifically, softmax requires computationally expensive division, is prone to numerical overflow, and has a sequential dependency on the fully computed matrix multiplication outputs, making it difficult to parallelize.

To address these challenges, we adopt the log-sum-exp transformation, which restructures the softmax computation into a sequence of simpler operations: finding the maximum value (y_{max}), computing the logarithm of summed exponentials, performing subtraction, and applying exponentiation. The transformation is expressed as:

$$Softmax(y_i) = \frac{\exp(y_i - y_{max})}{\sum_{j=1}^{D} \exp(y_j - y_{\max})}, \#$$
$$= \exp\left(y_i - y_{max} - \ln\left(\sum_{j=1}^{D} \exp(y_j - y_{max})\right)\right), \quad (6)$$

This decomposition allows for pipelined execution, significantly improving efficiency. As the Y matrix is generated from the preceding matrix multiplication (QK^T) in the scaled dot-product attention block, each element y_i is streamed directly into a 2-input 8-bit comparator that dynamically updates and stores y_{max} in local registers. Once y_{max} is available across all non-linear computation units, the subsequent log-sum-exp operations are executed in distinct processing stages, leveraging LUTs for logarithm and exponential functions while using dedicated adders/subtractors for intermediate calculations. By structuring the softmax function into pipelined modular operations, ASTRA reduces latency and hardware complexity while maximizing parallelism. Further details on data movement and pipelined execution are provided in Sections 3.6 and 3.7.

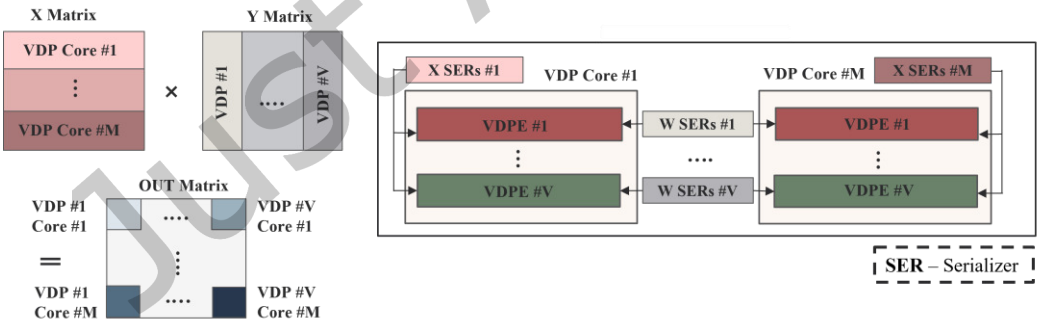


Figure 6: (a) Dataflow used in ASTRA where the X and W matrices in (a) are mapped onto VDPEs and VDP cores in (b).

3.6 Dataflow and Architecture-level Optimizations

ASTRA's dynamically-operated VDP cores enable flexible dataflow selection by dynamically encoding both operands. In contrast, most state-of-the-art photonic accelerators map one operand onto fixed photonic circuit states that cannot be readily reconfigured due to slow switching speeds, restricting these designs to a weight stationary (WS) dataflow only [10]–[15]. To efficiently execute GEMM operations in ASTRA, we

employ a fine-grained tiling strategy and meticulously designed spatial and temporal mappings, as shown in Figure 6.

While ASTRA adopts the widely used output stationary (OS) dataflow model—previously proposed in photonic accelerators such as [14]—our design introduces a hierarchical and architecture-aware adaptation that differs significantly in operand partitioning and hardware utilization. Specifically, ASTRA customizes OS dataflow to better align with the structure of its optical VDP architecture. Each VDP core handles the multiplication of one horizontally partitioned tile of matrix X with the full matrix W, similar to [14]. However, unlike [14], ASTRA further partitions W across the VDPEs within a core, enabling fine-grained operand distribution. This organization facilitates simplified control logic and efficient hardware sharing: the B_to_S and serializer circuits for matrix X are shared across all VDPEs within each VDP core, while those for matrix W are shared across identical VDPEs in different VDP cores, as illustrated in Figure 6(b). As a result, ASTRA's adaptation of OS dataflow not only matches the parallelism of the optical compute fabric but also reduces on-chip buffer requirements and power consumption.

As shown in Figure 6(a), 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. This partitioning scheme enables dynamic and fast operand switching through OSSMs and supports efficient sharing of B_to_S and serializer circuits, consistent with our architecture-aware OS dataflow mapping. Additionally, as discussed in Section 3.4, the PCAs enable temporal summation of a large number of multiplications, which reduces the required sampling rate for the ADCs in ASTRA to the MHz range. This lower sampling rate contributes to significant savings in both area and power for the ADCs, further optimizing the overall system efficiency.

3.7 Execution Pipelining and Scheduling

To fully exploit parallelism available in the photonic domain, ASTRA implements a pipelined execution model for transformer operations. Figure 7 illustrates the overall execution flow and the adopted pipelining strategy when accelerating an MHA layer. Based on Eq. (1), the MHA operations are divided into four main steps, as shown in the top-left section of Figure 7: (1) generating the Q, K, and V matrices, (2) computing the attention score $Q \times K^T$, (3) performing softmax, and (4) generating the final MHA output $S \times V$. Other than the softmax operation, these steps primarily consist of matrix multiplications, following the same general execution sequence in a non-pipelined approach.

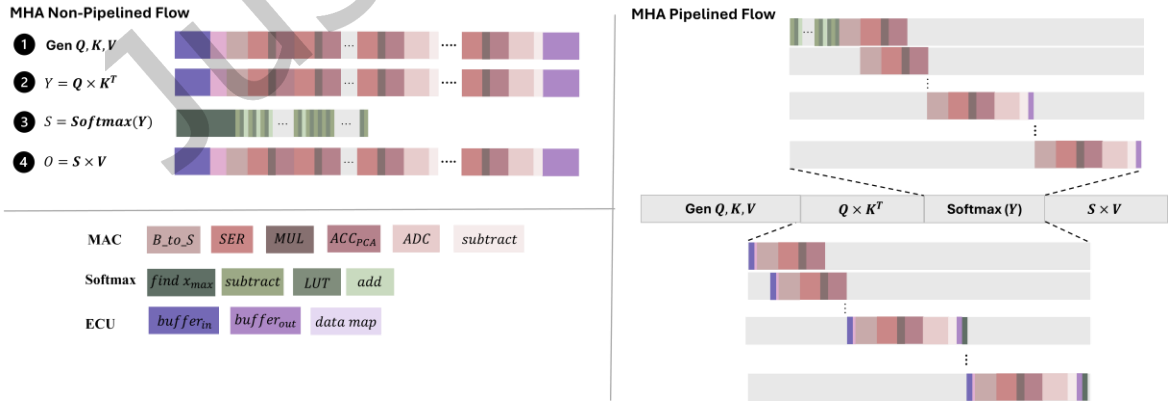


Figure 7: ASTRA's MHA layer pipelining and execution flow.

In a conventional non-pipelined execution, each matrix multiplication begins by fetching operand matrices from the ECU buffers. The parameters are then mapped to the B_to_S and serializer units based on the dataflow model described in Section 3.6. The serializers sequentially feed stochastic bit streams to the

OSSMs, where the multiplications occur. The resulting optical signals propagate through waveguides to the PCA, where they are accumulated. Once an output value is ready, it undergoes analog-to-digital conversion (using the ADCs) before subtraction. Finally, the processed value is stored in the post-processing unit within the ECU. However, as depicted in Figure 7, this non-pipelined execution forces sequential waiting for each time-consuming step—including B_to_S conversion, serialization, PCA accumulation, and ADC conversion—leading to inefficiencies, especially given the massive computational requirements of transformer models. Furthermore, interfacing with main memory becomes a major bottleneck, as the frequent retrieval of large data chunks negates the advantages of the OS dataflow outlined in Section 3.6, leading to increased energy inefficiencies.

To address this, ASTRA efficiently partitions these operations, enabling a highly optimized pipelining model. During matrix multiplications, ASTRA concurrently executes: (i) data fetching and mapping, (ii) B_to_S data conversion, (iii) bit-stream generation via serializers, (iv) optical AND operations in the OSSMs, (v) bit accumulation in the PCA, (v) ADC conversion and subtraction, and (vii) data storage in ECU buffers to be used in the subsequent computations. This approach effectively overlaps latencies associated with each computational step, as illustrated in Figure 7, significantly improving overall efficiency. The pipelining model is applied to all matrix multiplication operations within the MHA and FFN layers across both transformer encoder and decoder blocks. To further minimize latency, ASTRA integrates softmax optimizations into its pipelining scheme. Specifically, as attention score matrices are generated within each VDPE, their output values are concurrently processed by softmax 8-bit comparators, which continuously update y_{max} (see Eq. (5)). Additional softmax computations, such as subtractions and exponential calculations, are also pipelined alongside the $S \times V$ computation, as shown in Figure 7.

4 EVALUATION

4.1 Simulation Setup

We conducted comprehensive device- and architecture-level simulation-based analyses to evaluate the efficiency of the ASTRA architecture. Five transformer neural network models were considered in the experiments: Transformer-base, BERT-base, Albert-base, ViT-base, and OPT-350. The model parameters for these are shown in Table 1. A Python-based simulator was developed to estimate the performance and energy costs for running each model. The simulator accounts for both software and hardware mapping, performing layer-wise mapping for each transformer model and dataset while accurately modeling all peripherals and devices, as detailed in Table 2. Various factors were considered for assessing photonic signal losses and power, including waveguide propagation loss (1 dB/cm [13]), splitter loss (0.13 dB [38]), combiner loss (0.9 dB [38]), MR through loss (0.02 dB [15]), and OAG tuning and control power (6mW) [39], and 3.5dB insertion loss. The comb laser source from [40] was used with >-3 dBm optical power output at 25 usable wavelengths with 0.5W wall-plug power. Performance and energy estimates for the LUTs and electronic buffers in ASTRA were derived using CACTI [41]. The electronic circuits for softmax and B_to_S converters were synthesized using Xilinx Vivado. Model training and accuracies were evaluated using PyTorch 2.3.

Our analysis with transformer model precision shows that using 8-bit quantization for models results in transformer inference accuracy comparable to that achieved with full precision (FP32), as shown in Table 3. The % accuracy metric is used to evaluate the performance of transformer-base, BERT-base, Albert-base, and ViT-base models applied to tasks such as translation, sentiment analysis, and image classification. The BLEU score is reported for the OPT-350 model, which is used for text generation. Based on this evaluation, we have selected transformer models with 8-bit precision, where ASTRA represents each model parameter stochastically with 128 bits plus one sign bit.

Table 1: Transformer Models Configuration

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 (ns)	Power (mW)	Area (mm^2)
Comparator	0.6237	0.055	8.8E-09
Adder/Subtractor	0.7199	0.0028	5.5E-09
LUT	0.2225	1.403	1.597E-06
B_to_S	0.5302	0.021	6.3E-08
Serializer [42]	0.03	1.5	0.0021
ADC [43]	0.78	2.55	0.002
PCA	0.033~2.19	0.02	0.28
Attenuators in OAGs [25]	~0.01	0.00001	0.00002
OSSMs	~0.01	~1	0.0001

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)

4.2 Device-level Scalability and Error Analysis

We operate our OAGs at 30 Gbps and set the sampling bandwidth of PDs to 230MHz. This enables us to have $\beta = 2^B - 1 = 128$ since $30\,GHz \div 230\,MHz > 128$. At 230MHz sampling bandwidth and thermal noise floor, the PD/BPD can perform incoherent superposition of 0.2μW optical pulses [30]. Therefore, by accounting for the conservative insertion loss of 4dB for the OAG and incorporating additional 7dB losses for optical power (due to coupling and propagation losses, etc.) [31], [32] to reach the OAG from the laser source, each OAG requires 0.5μW optical power to operate. This means that if a comb laser source provides ~512μW (−3dBm) optical power per wavelength, a total of ~1024 OAGs can be supported per wavelength. Therefore, homodyne VDPEs, each comprising ~1024 OAGs, can be employed in ASTRA. This enables massive processing parallelism at low optical power.

To verify the achievable OAG speed and PCA capacity (i.e., number of streaming stochastic X and W values that can be multiplied and accumulated), we modeled and analyzed our VDPE design using Lumerical Interconnect and Cadence Virtuoso. The results are shown in Figure 8. Evidently, for the input wavelength power of 0.5mW and output pulse power of -37dBm (~0.2μW), a total of 1024 OAGs (actually up to 1027 to be more precise) can be supported at the photon lifetime limited speed of >30Gbps (Figure 8(a)). Moreover, at 30Gbps OAG speed, the PCA capacity is >10,000 (Figure 8(b)).

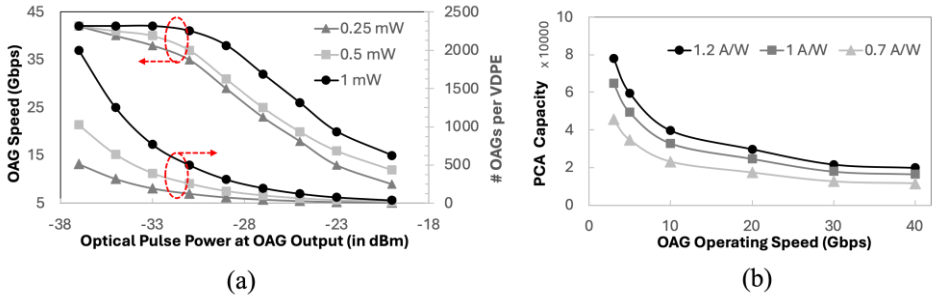


Figure 8: VDPE scalability results. (a) Achievable OAG speed and #OAGs per VDPE versus optical pulse power at OAG output for different input wavelength power. (b) PCA capacity versus OAG speed for different PD responsivities.

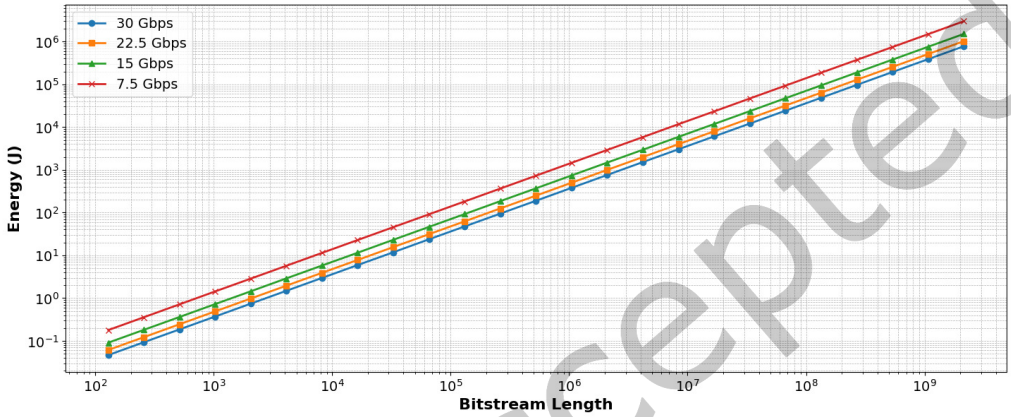


Figure 9: ASTRA's average inference energy consumption across five transformer models (Table 1) versus bitstream length for different OAG operating speeds.

Building on this analysis, we conducted a comprehensive error evaluation of our OSSM using a methodology like that employed in prior studies [12], [23]. First, we determined the mean absolute error of our OSSM design through detailed simulations. This error was then introduced into the inference process of various transformer models using PyTorch. By avoiding stochastic additions and leveraging the low-error, low-cost multiplication technique discussed in Section 3, our OSSM achieved a mean absolute error of 0.042 for signed multiplications, surpassing the accuracy of stochastic multipliers reported in previous studies [12]. When applied to transformer model inference, ASTRA exhibited minimal accuracy degradation, as shown in Table 3 (Q(8-bit) + SC), averaging 1.15% (0.58 BLEU) compared to FP32 and 0.25% (0.3 BLEU) compared to quantized 8-bit models.

Furthermore, we conducted a bitstream scalability analysis to evaluate how the average inference energy consumption across the five transformer models from Table 1, scales with increasing bitstream lengths, as shown in Figure 9. The results exhibit a near-linear trend on a log-log scale, reflecting the underlying exponential growth in energy as bitstream length increases. Specifically, as the bitstream length grows from 129 bits (corresponding to 8-bit binary precision) to over 2.1×10^9 bits (corresponding to 32-bit binary precision), latency and energy rise sharply. Even a single-bit increase in binary precision (e.g., from 16 to 17 bits) doubles the bitstream length and results in a substantial jump. In addition, Figure 9 shows four distinct lines corresponding to different OAG operating speeds, demonstrating how energy varies across configurations. While higher-speed OAGs (e.g., 30 Gbps) incur slightly higher instantaneous power due to increased switching activity, they benefit from significantly shorter accumulation times. Conversely, lower-speed OAGs (e.g., 7.5 Gbps) lead to prolonged execution, which dominates total energy consumption despite slightly lower power. These findings highlight that ASTRA's default configuration of 128-bit stochastic

operands offers a well-balanced trade-off among latency, energy, and inference accuracy, as supported by the quantization error analysis in Table 3. Overall, the analysis underscores the importance of jointly optimizing bitstream length and OAG speed to ensure scalable and energy-efficient operation within ASTRA's precision-tunable architecture.

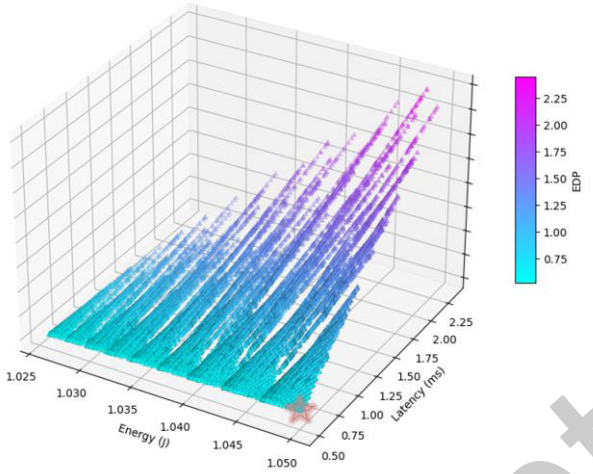


Figure 10: Architectural design-space exploration for ASTRA, to find the optimal $[M, V, N]$ configuration with the best EDP. The best configuration, [106, 25, 515] is shown with the pink star.

4.3 Architecture design space exploration

The ASTRA architecture relies on three main parameters, as outlined in Section 3: $M, V,$ and N. M represents the number of VDP cores, explored within the range $[1,200]$; V denotes the number of VDPEs per VDP core, within $[1,25]$; and N specifies the number of OAGs per VDPE, ranging from $[1,1024]$. To determine the optimal configuration for ASTRA—defined as the combination of $[M, V, N]$ that yields the lowest energy-delay product (EDP)—we performed an exhaustive design space exploration. Given the vast number of generated configurations (over 5 million data points), Figure 10 presents a sampled subset of the design space for improved visibility and representation. Using the ASTRA simulator discussed in Section 4.1, the energy and latency values were obtained for each transformer model and each accompanying dataset for the wide set of possible values for $[M, V, N]$. The average EDP values across all the transformer models and datasets for each set of parameters were then obtained and the optimal configuration $[106, 25, 515]$ was identified as the one with the lowest EDP values.

4.4 Architectural Component-Wise Energy Analysis

To analyze the energy overheads of key components within the ASTRA architecture, we present an energy breakdown in Figure 11. Notably, the serializers and OAGs account for more than half of the total energy overhead. This is primarily due to the large matrix dimensions in transformer layers, where multiplication operations are governed by the OSSMs. As discussed in Section 4.2, ASTRA operates OAGs and serializers at speeds of up to 30 Gbps, enabling fast, seamless, and continuous multiplications via stochastic bitwise AND operations. Additionally, the use of a comb laser per each VDP core reduces power consumption compared to traditional laser sources, enhancing overall energy efficiency in optical accelerators. While DACs and ADCs are typically among the most energy-intensive components in optical accelerators, ASTRA minimizes their energy impact using our proposed OSSM designs. Although ADCs remain necessary, their use is limited to the final output stage before buffering, as all accumulations are performed in-situ using the PCAs.

Moreover, ASTRA minimizes the overhead of non-linear functions by implementing activation functions via LUTs and simple digital circuits, along with the optimized and pipelined softmax computation. However,

as illustrated in Figure 11, the softmax operation becomes more computationally intensive for larger models such as OPT-350, due to the increased number of values that must be processed following the generation of attention scores in the MHA layers. Lastly, ALBERT achieves additional energy savings by sharing attention and feedforward parameters across layers [4], reducing the number of active serializers and consequently lowering overall energy consumption.

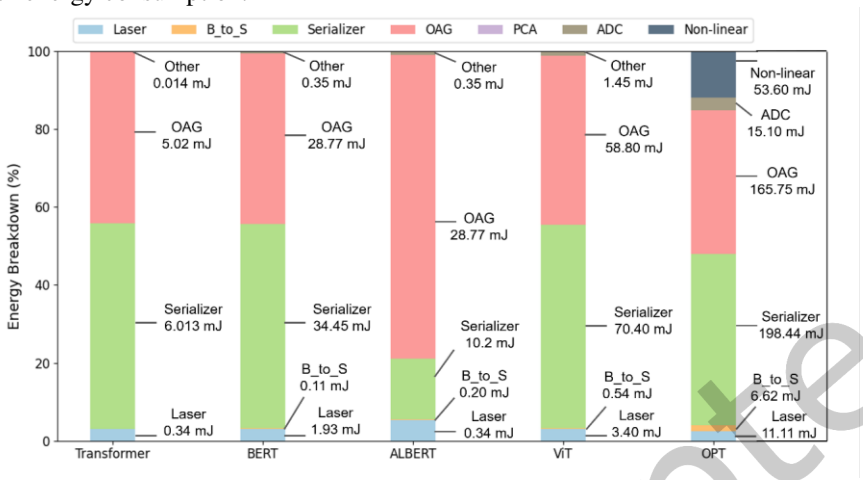


Figure 11: Energy breakdown across ASTRA components.

4.5 Comparison with State-of-the-art Hardware Accelerators

ASTRA is compared against CPU, GPU, TPU, and several state-of-the-art transformer accelerators: an FPGA-based transformer accelerator FPGA_ACC [5], a processing-in-memory accelerator, TransPIM [7], an MZM-based optical accelerator Lightning-Transformer (LT) [14], an MR-based optical accelerator TRON [15],and an optical accelerator that leverages stochastic computing SCONNA [20].

4.5.1 Speedup Comparison

Figure 12 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×, 7.6×, and 5.1× over CPU, GPU, TPU, FPGA_ACC, TransPIM, TRON, LT, and SCONNA, respectively. As illustrated in the figure, while FPGA_ACC outperforms CPU, GPU, and TPU, the superior performance of TransPIM, LT, TRON, SCONNA and ASTRA highlights the limitations of electronic accelerators in meeting the computational demands of transformer neural networks. Moreover, the higher performance of LT, TRON, SCONNA, and ASTRA compared to TransPIM underscores the growing potential of silicon photonics and its ability to surpass in-memory computing. The significantly lower latencies achieved by ASTRA can be attributed to its high degree of parallelism, enabled by the proposed VDP core design, as well as the integration of device-, circuit-, and architectural-level optimizations, and the adoption of stochastic computing. While SCONNA also leverages stochastic computing, its scalability and parallelism are constrained by heterodyne crosstalk in its WDM-based cascaded MR design, which limits the number of MRs per VDPE. In contrast, ASTRA's novel homodyne VDPE architecture effectively mitigates these limitations, enabling significantly improved performance.

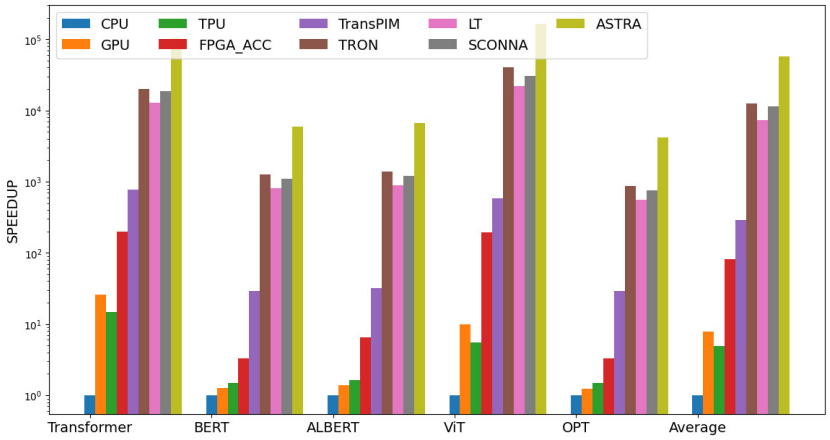


Figure 12: Speedup comparison for CPU, GPU, TPU, transformer accelerators (FPGA_ACC [5], TransPIM [7], LT [14], TRON [15], SCONNA [20]), and ASTRA.

4.5.2 Energy Efficiency Comparison

The energy comparison results for ASTRA against the compute platforms and transformer accelerators are presented in Figure 13, 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×, 1.3×, and 1.6× compared to the CPU, GPU, TPU, FPGA_ACC, TransPIM, TRON, LT, and SCONNA, respectively. The lower energy consumption of TransPIM compared to CPU, GPU, TPU, and FPGA_ACC demonstrates the advantages of in-memory computing in reducing data movement overheads. However, the higher energy efficiency of LT, TRON, SCONNA, and ASTRA over TransPIM further underscores the potential of silicon photonics, which exploits low-power, high-speed optical operations. SCONNA, in particular, demonstrates competitive energy efficiency owing to its use of stochastic computing. Nevertheless, its energy gains are tempered by the need for additional peripheral circuitry, such as LUTs and serializers for stochastic bitstream generation, and by architectural limitations related to WDM-based cascaded MRs. In contrast, ASTRA surpasses all evaluated accelerators in energy efficiency due to its highly optimized VDPE architecture, the lightweight OSSM design, the elimination of power-hungry DACs, and its extremely low-latency optical operation.

4.6 Precision Scalability and Comparison

Recent research has advocated for model compression techniques, particularly through quantization, where model precision is reduced to below 8-bits to improve energy efficiency while maintaining high accuracy [44]. We conducted a precision scalability analysis to assess the energy consumption of ASTRA when running transformer models with 4-bit quantization (ASTRA-4bit), compared to our baseline architecture using 8-bit quantization (ASTRA-8bit). SC requires 2^N bits for each N-bit binary number, and thus reducing bit-width, can lead to improvements in both performance and energy efficiency.

In contrast, other hardware accelerators, such as LT [14], which is the only other accelerator that also evaluates 4-bit transformer models, show more modest energy savings when reducing precision from 8-bits (LT-8bit) to 4-bits (LT-4bit). Although LT benefits from an exponential reduction in analog amplitude levels (e.g., from 256 to 16 levels), these symbols still require high optical power to ensure error-free transmission and detection. Meanwhile, ASTRA operates using binary stochastic bitstreams, which only require two amplitude levels (0 and 1), allowing significantly lower optical power per operand. As a result, ASTRA achieves more pronounced energy gains with quantization due to both the exponential reduction in bitstream length and the inherently lower power required for binary optical modulation.

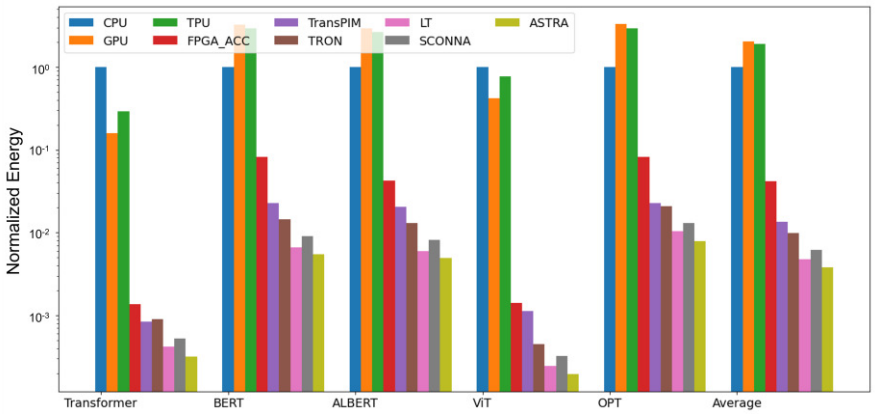


Figure 13: Energy comparison for CPU, GPU, TPU, transformer accelerators (FPGA_ACC [5], TransPIM [7], LT [14], TRON [15], SCONNA [20]), and ASTRA.

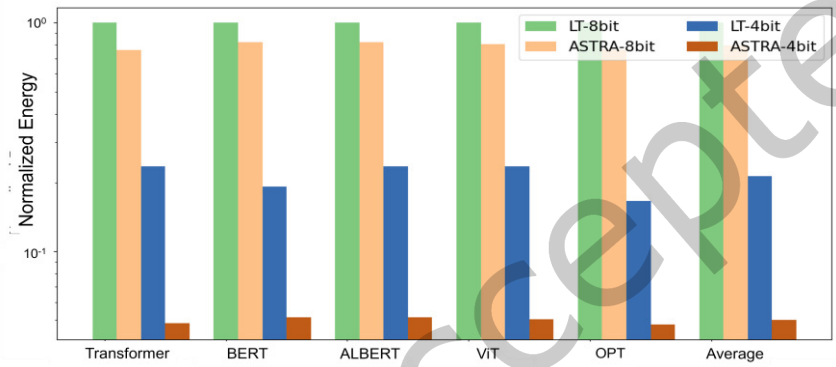


Figure 14: Energy comparison between 8-bit and 4-bit variant models on LT [14] and ASTRA optical transformer accelerators.

Figure 14 shows the energy consumption of both LT and ASTRA's 8-bit and 4-bit variants, normalized to LT-8bit results for each transformer model. On average, ASTRA-4bit achieves a 20.8×, 16.0×, and 3.5× reduction in energy consumption compared to LT-8bit, ASTRA-8bit, and LT-4bit, respectively. These results highlight that while ASTRA-8bit outperforms all the transformer accelerators explored, further bit-width reduction can lead to even greater energy savings.

4.7 Physical Implementation Feasibility and Area Analysis

Figure 15 presents the area breakdown of our proposed architecture, which occupies a total area of 295.75 mm². The largest contributors to this area are the PCAs and OAGs, accounting for 50.18% and 46.15% of the total footprint, respectively. Additional components include serializers, ADCs, B_to_S converters, and non-linear units. Leveraging the dataflow architectural optimization described in Section 3.6, ASTRA significantly reduces area overhead through the efficient sharing of serializers and B_to_S converters across OSSMs. Furthermore, the use of comb lasers—requiring only one laser per VDP core—substantially minimizes the area otherwise consumed by discrete laser sources. The non-linear units, comprising softmax and activation functions, contribute minimally to the overall area due to our compact implementation using LUTs and optimized circuit designs. These optimizations collectively enable ASTRA to integrate a very high degree of parallelism, supporting a total of 1,364,750 OSSMs (computed as 515 OSSMs × 25 VDPEs × 106 VDP cores). While ASTRA's total area is moderately higher than that of some prior optical accelerators such as [14], its capacity to support approximately 1.4 million concurrent MACs results in superior area efficiency. Specifically, ASTRA achieves 0.2166 mm² per 1,000 parallel MACs, demonstrating a favorable compute-to-area ratio compared to state-of-the-art accelerators.

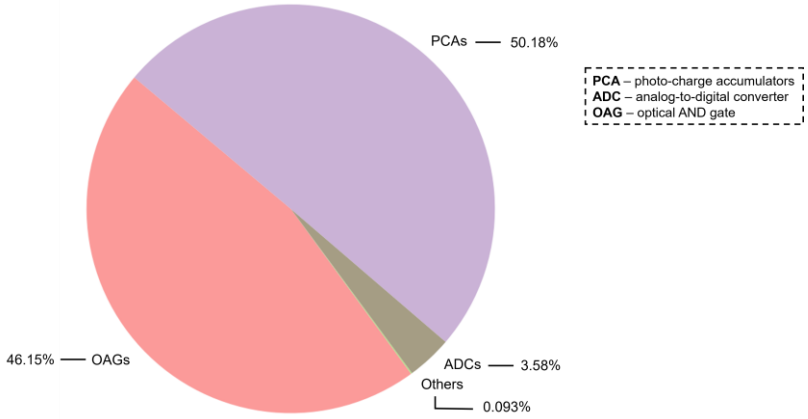


Figure 15: Area breakdown across ASTRA components.

ASTRA can be fabricated using a monolithic silicon photonic process or through 3D integration, where photonic layers are stacked above underlying electronic control layers. Both approaches align with recent trends in heterogeneous integration, and advances in wafer-scale silicon photonics have demonstrated the viability of integrating tens of thousands of MRs, PDs, and related devices on a single die [33]. The feasibility of such high-density integration has been validated in prior photonic neural network implementations, which employ scalable architectures based on MR banks, shared waveguides, and compact photonic layouts [10]-[16]. For example, designs utilizing WDM with MR weight banks enable high fan-in connectivity within a compact footprint. Architectures such as Broadcast-and-Weight and Modulator Neuron further illustrate the effectiveness of integrating large numbers of passive photonic components and detectors using CMOS-compatible fabrication [47]. Techniques including wavelength reuse, waveguide sharing, and layout-aware design optimization collectively support ASTRA's scalability and establish its practical feasibility for chip-level integration.

5 CONCLUSION

In this paper, we introduced ASTRA, a novel stochastic optical accelerator designed for transformer neural networks. ASTRA integrates stochastic, optical, and analog computing while advancing the capabilities of optical VDP cores with several design innovations. The proposed optical homodyne VDPEs, which integrated stochastic signed multipliers, achieved significantly reduced latency and energy consumption by leveraging stochastic computing for multiplications and analog-domain accumulations. Compared to GPU, TPU, CPU, and several state-of-the-art transformer neural network accelerators, ASTRA 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 transformer neural networks efficiently.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS, 2017.*

[2] J. Devlin, M. W. Chang, K. Lee, K., and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR, Oct 2018.*

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, and R. Avila, "Gpt-4 technical report," arXiv preprint arXiv:2303.08774., 2023.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR, Oct. 2020.*

[5] S. Lu, M. Wang, S. Liang, J. Lin, and Z. Wang, "Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer," *IEEE SOCC, 2020.*

[6] P. Qi, E. H. M. Sha, Q. Zhuge, H. Peng, S. Huang, Z. Kong, Y. Song, and B. Li, "Accelerating framework of transformer by hardware design and model compression co-optimization," *IEEE ICCAD, 2021.*

[7] M. Zhou, W. Xu, J. Kang and T. Rosing, "TransPIM: A Memory-based Acceleration via Software-Hardware Co-Design for Transformer," *IEEE HPCA, 2022.*

[8] L. Siyuan, M. Wang, S. Liang, J. Lin, and Z. Wang , "Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer," *IEEE SOCC, 2020.*

[9] C. E. Leiserson, N. C. Thompson, J. S. Emer, B. C. Kuszmaul, B. W. Lampson, D. Sanchez, and T. B. Schardl, "There's plenty of room at the Top: What will drive computer performance after Moore's law?" *Science 368.6495, 2020.*

[10] F. Sunny, A. Mirza, M. Nikdast, and S. Pasricha, "CrossLight: A cross-layer optimized silicon photonic neural network accelerator." *ACM/IEEE DAC,2021.*

[11] V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. De Lima, H. T. Peng, P. R. Prucnal, and B. J. Shastri, "Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs)," *IEEE JQE, 2020.*

[12] F. Sunny, M. Nikdast, and S. Pasricha, "RecLight: A Recurrent Neural Network Accelerator with Integrated Silicon Photonics." *IEEE ISVLSI, 2022*

[13] S. Afifi, F. Sunny, A. Shafiee, M. Nikdast, and S. Pasricha, "GHOST: A Graph Neural Network Accelerator using Silicon Photonics." *ACM TECS, 2023*

[14] 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." *IEEE HPCA, 2024*.

[15] S. Afifi, F. Sunny, M. Nikdast, and S. Pasricha, "Tron: Transformer neural network acceleration with non-coherent silicon photonics." *GLSVLSI, 2023.*

[16] S. Afifi, I. Thakkar, and S. Pasricha, "SafeLight: Enhancing Security in Optical Convolutional Neural Network Accelerators", *DATE, 2025.*

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut , "Albert: A lite bert for self-supervised learning of language representations," *ICLR, Sep. 2019.*

[18] S. Li, A. O. Glova, X. Hu, P. Gu, D. Niu, K. T. Malladi, H. Zheng, B. Brennan, and Y. Xie, "Scope: A stochastic computing engine for dram-based in-situ accelerator," *IEEE/ACM MICRO, 2018.*

[19] S. Afifi, I. Thakkar, and S. Pasricha, "ARTEMIS: A Mixed Analog-Stochastic In-DRAM Accelerator for Transformer Neural Networks," *IEEE/ACM CASES (ESWEEK), Oct 2024.*

[20] S. S. Vatsavai, V. S. P. Karempudi, I. Thakkar, A. Salehi, and T. Hastings, "SCONNA: A Stochastic Computing Based Optical Accelerator for Ultra-Fast, Energy-Efficient Inference of Integer-Quantized CNNs," *IEEE IPDPS*, *2023*.

[21] D. Wu, J. Li, R. Yin, H. Hsiao, Y. Kim, and J. San Miguel, "UGEMM: Unary Computing Architecture for GEMM Applications," *ACM/IEEE ISCA*, *2020.*

[22] S. Afifi, F. Sunny, M. Nikdast, and S. Pasricha, "Accelerating Neural Networks for Large Language Models and Graph Processing with Silicon Photonics", *IEEE/ACM DATE, 2024.*

[23] S. S. Vatsavai and I. Thakkar, "A Bit-Parallel Deterministic Stochastic Multiplier," *ISQED, 2023.*

[24] S. M. Shivanandamurthy, I. G. Thakkar, and S. A. Salehi, "Atria: A bit-parallel stochastic arithmetic based accelerator for in-dram cnn processing." *IEEE ISVLSI, 2021.*

[25] C. Ye, S. Khan, Z. R. Li, E. Simsek, and V. J. Sorger, "λ-Size ITO and Graphene-Based Electro-Optic Modulators on SOI," *IEEE Journal of Selected Topics in Quantum Electronics, 2014.*

[26] J. K. George *et al.*, "Neuromorphic photonics with electro-absorption modulators," *Opt. Express,* 2019.

[27] "Pic design and simulation software- lumerical interconnect," Apr 2021. [Online]. Available: https://www.lumerical.com/products/interconnect/

[28] F. Brückerhoff-Plückelmann, I. Bente, D. Wendland, J. Feldmann, C. D. Wright, H. Bhaskaran, and W. Pernice, "A large scale photonic matrix processor enabled by charge accumulation," *Nanophotonics*, *2023.*

[29] S. S. Vatsavai, V. S. P. Karempudi, and I. Thakkar, "An Optical XNOR-Bitcount Based Accelerator for Efficient Inference of Binary Neural Networks," *ISQED*, *2023.*

[30] A. Sludds, et al., "Delocalized photonic deep learning on the inter net's edge," *Science*, *2022.*

[31] M. A. Al-Qadasi, L. Chrostowski, B. J. Shastri, and S. Shekhar, "Scaling up silicon photonic-based accelerators: Challenges and opportunities," *APL Photonics*, *2022*.

[32] S. S. Vatsavai, I. G. Thakkar, "Photonic Reconfigurable Accelerators for Efficient Inference of CNNs With Mixed-Sized Tensors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *2022*.

[33] Lightelligence, [Online]: https://www.lightelligence.ai/, Accessed on: Nov 15, 2024.

[34] Lightmatter, [Online]: https://lightmatter.co, Accessed on: Nov 15, 2024.

[35] Luminous, [Online]: https://www.luminous.com, Accessed on: Nov 15, 2024.

[36] Cognifiber, [Online]: https://www.cognifiber.com/, Accessed on: Nov 15, 2024.

[37] T. Ferreira de Lima, E. A. Doris, S. Bilodeau, W. Zhang, A. Jha, H. T. Peng, E. C. Blow, C. Huang, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "Design automation of photonic resonator weights," *Nanophotonics,* 2022.

[38] L. H. Frandsen, P. I. Borel, Y. X. Zhuang, A. Harpøth, M. Thorhauge, M. Kristensen, W. Bogaerts, P. Dumon, R. Baets, V. Wiaux, and J. Wouters, "Ultralow-loss 3-dB photonic crystal waveguide splitter," *Optics letters,* 2004.

[39] 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, 2023.*

[40] A. Rizzo, A. Novick, V. Gopal, B. Y. Kim, X. Ji, S. Daudlin, Y. Okawachi, Q. Cheng, M. Lipson, A. L. Gaeta, and K. Bergman, “Massively scalable Kerr comb-driven silicon photonic link,” *Nat. Photon.*, *2023.*

[41] HP Labs : CACTI. [Online]: https://www.hpl.hp.com/research/cacti/.

[42] S. Lin, S. Moazeni, K. T. Settaluri, and V. Stojanović, "Electronic–Photonic Co-Optimization of High-Speed Silicon Photonic Transmitters," *Journal of Lightwave Technology*, *2017.*

[43] D. R. Oh, K. J. Moon, W. M. Lim, Y. D. Kim, E. J. An, and S. T. Ryu, “An 8b 1gs/s 2.55mw sar-flash adc with complementary dynamic amplifiers,” *IVLSIC, 2020.*

[44] H. Xi, C. Li, J. Chen, and J. Zhu, "Training transformers with 4-bit integers." *Advances in Neural Information Processing Systems*, *2023.*

[45] Y. Ding, C. Liu, M. Duan, W. Chang, K. Li, and K. Li, "HAIMA: A Hybrid SRAM and DRAM Accelerator-in Memory Architecture for Transformer." *ACM/IEEE Design Automation Conference (DAC)*, 2023.

[46] F. Sunny, E. Taheri, M. Nikdast, and S. Pasricha “A survey on silicon photonics for deep learning.” *ACM Journal of Emerging Technologies in Computing System*, 2021.