NOCTUA: A High-Efficiency Reconfigurable NoC-based Transformer Universal Accelerator

Kun-Chih (Jimmy) Chen, Pin-Ching Shen, and Bo-Chun Chen
Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
Email: kcchen@nycu.edu.tw

***Abstract*—Transformer models, particularly Large Language Models (LLMs), are revolutionizing numerous fields but impose significant computational demands, necessitating specialized hardware acceleration. This paper introduces NOCTUA, a novel Network-On-Chip (NoC)-based Transformer Universal Accelerator, designed and presented as a High-Efficiency Reconfigurable NoC-based Transformer Universal Accelerator. NOCTUA leverages a scalable Network-on-Chip, renowned for its flexibility and high-bandwidth communication, to interconnect a multitude of Processing Elements (PEs). These PEs are designed to execute complex Transformer operations both independently and in coordinated concert with other PEs, enabling fine-grained parallelism and collaborative computation. This distributed and reconfigurable design, underpinned by the adaptable NoC backbone, allows NOCTUA to dynamically adapt its hardware resources and dataflow pathways, achieving both high operational efficiency and universal applicability across a wide spectrum of Transformer models and workloads. We present the architectural details of NOCTUA and demonstrate its significant potential to enhance processing throughput and energy efficiency compared to conventional approaches. Through experiments with existing attention-based NLP models, including BERT and GPT-2 on various language tasks, NOCTUA achieves a performance of 1.05 TOPS/W and outperforms these models in area efficiency with 0.71 TOPS/mm^2.**

***Index Terms*—Network-on-chip, flexible dataflow processing, neural network accelerators, transformers**

I. Introduction

The Transformer architecture has become fundamental to modern Artificial Intelligence (AI) models, driving significant advancements in natural language processing, computer vision, and various other applications [1]–[3]. The success of these models is primarily due to their powerful self-attention mechanism, which effectively captures long-range dependencies within input sequences. However, this mechanism often results in quadratic complexity related to the sequence length, leading to a substantial increase in computational requirements for longer inputs. Consequently, this impressive performance comes with considerable computational demands and large model parameters, with state-of-the-art models now featuring billions of parameters and requiring significant memory resources. As model sizes continue to grow, the need for high-performance hardware accelerators that can efficiently handle these complexities has become increasingly urgent [6].

Due to the demands of edge computing, various hardware architectures for Transformers have been proposed that aim to enhance efficiency for specific Transformer models by customizing the hardware to suit the algorithm [8], [9]. However, these architectures often lack the flexibility needed to adapt to rapidly evolving model designs and diverse application scenarios. This inflexibility, combined with high non-recurring engineering costs and long design cycles, makes these ASICs vulnerable to becoming obsolete quickly as new variants of Transformers emerge. Additionally, they face challenges in meeting the varying computational demands of different large language model (LLM) matrix operations, where fixed processing element (PE) arrays can suffer from underutilization. This leads to wasted resources and decreased energy efficiency. In contrast, more flexible network-on-chip (NoC)-based approaches can better address these issues, as illustrated in Fig. 1.

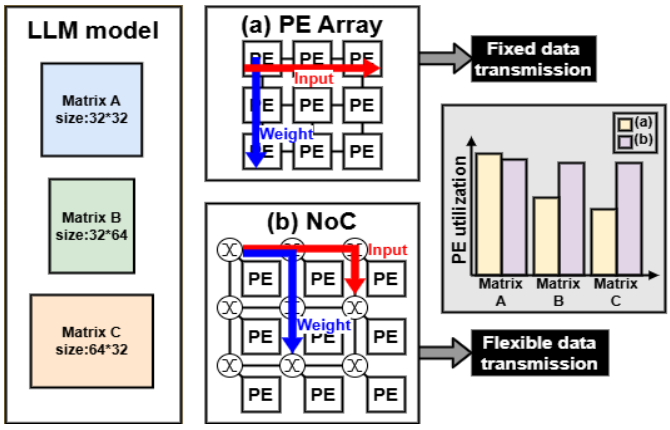

Fig. 1. (a) The conventional PE array accelerator and (b) the proposed NoC-based accelerator apply with LLM models, and the proposed method outperforms at PE utilization.

To address the challenges in deep learning acceleration, we propose the NoC-based Transformer Universal Accelerator (NOCTUA), a highly efficient and dataflow-reconfigurable solution designed specifically for Transformers. Utilizing a Network-on-Chip (NoC) [10] as its scalable communication backbone, NOCTUA effectively manages communication among numerous specialized Processing Elements (PEs), which are tailored for Transformer operations. This structure promotes high levels of parallelism and modular design while minimizing data movement overhead, a critical bottleneck in acceleration, thus enhancing operational efficiency. Furthermore, NOCTUA's ability to dynamically adapt dataflow pathways and processing methods among PEs based on the sizes and characteristics of matrix operations in various Trans-

979-8-3315-7206-8/25/$31.00

former models counters the rigidity of traditional architectures, enabling effective support for diverse Transformer variants and their configurations, ultimately achieving universal acceleration. The main contributions of this paper are as follows:

- We involve a flexible NoC interconnection to propose a NOCTUA, which supports a reconfigurable dataflow mapping mechanism, and thereby achieving high throughput and energy efficiency.
- We propose a novel layer normalization, non-linear activation functions, and softmax methods, which are common core operations in the Transformer model, to improve computing efficiency.

In this work, we evaluate the proposed NOCTUA by computing various attention-based NLP models, such as BERT and GPT-2. The experimental results show that the proposed NOCTUA can achieve area efficiency by 0.71 TOPS/mm^2 and energy efficiency by 1.05 TOPS/W.

II. Related Work

Some accelerators are designed for specific platforms, such as the reconfigurable RAWAtten [12] or the FPGA-based Unified Accelerator [13]. However, these designs have inherent limitations. For example, RAWAtten's centralized, bus-based interconnect can create a bottleneck in scalability. Additionally, unified designs for both attention and convolution often result in suboptimal data paths. In contrast, NOCTUA, which is a dedicated ASIC, utilizes a scalable Network-on-Chip (NoC) to effectively avoid these contention issues.

Other designs, such as ITA [11], optimize for a specific data flow. Although ITA's fixed weight-stationary approach is efficient for certain operations, this rigidity can result in underutilization of processing elements (PEs) when dealing with the varying matrix sizes typical of Transformer models. In contrast, NOCTUA addresses this issue by utilizing its Network-on-Chip (NoC) to implement a reconfigurable data flow, allowing it to adjust data paths dynamically for each layer. Additionally, the NoC incorporates efficient multicast capabilities, which is a significant advantage for the one-to-many data distribution required in Transformers, effectively reducing both latency and memory traffic.

Regarding nonlinear functions, ITA [11] introduces a new integer-only Softmax. However, scaling-based methods can be sensitive to input distributions, especially in unpadded sequences where attention scores may be very close in value. NOCTUA addresses this issue by providing a dynamic range-adaptive integer Softmax. This method adjusts its internal scaling factor based on the dynamic range of the input, which helps maintain sharp attention focus even without padding. As a result, it offers a more robust hardware solution for masked attention.

III. NoC-based Transformer Universal Accelerator (NOCTUA) Processor Design

A. NoC Architecture and Reconfigurable Dataflow

The proposed NOCTUA involves the Network-on-Chip (NoC) interconnection to support reconfigurable dataflow. This allows the accelerator to dynamically adapt data paths and distribution methods between its Processing Elements (PEs) to optimally handle the specific characteristics of any given operation, such as large-scale matrix multiplications, complex self-attention interactions, or varied matrix sizes. Unlike traditional NoCs, NOCTUA integrates hardware-efficient multicast mechanisms, enabling shared data, including weights or input, to be distributed to multiple PEs in a single operation. In this way, the proposed NOCTUA is designed to reduce the latency of data delivery and minimize memory access. Furthermore, a dedicated controller performs data tiling to partition large matrices into smaller blocks, ensuring data is perfectly sized for PE processing and optimized for exchange over the NoC.

Fig. 2 shows the architecture of the proposed NOCTUA, which is controlled by the centralized controller. In addition to the 4x4 2D mesh for the interconnection between each PE, the two on-chip memory units are used to optimize data access. To balance computational capability with hardware resources, NOCTUA employs a heterogeneous PE design. As shown in Fig. 3, PEs in the leftmost column integrate dedicated nonlinear units alongside their systolic arrays and

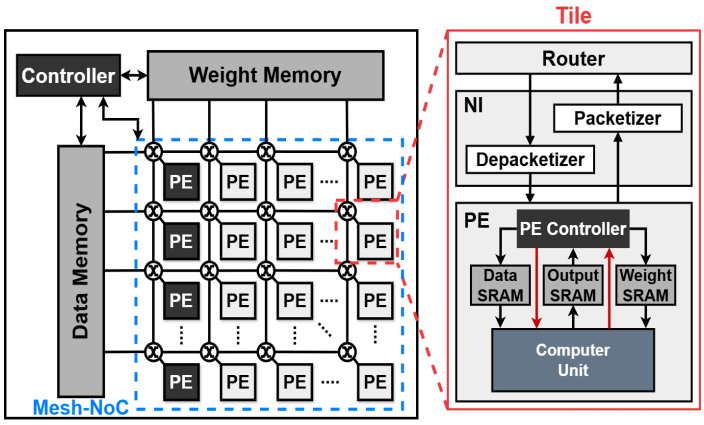

Fig. 2. NOCTUA accelerator architecture, showing the central controller, Mesh-NoC connecting PEs and memories, and a Tile inset detailing its router, NI, and PE components.

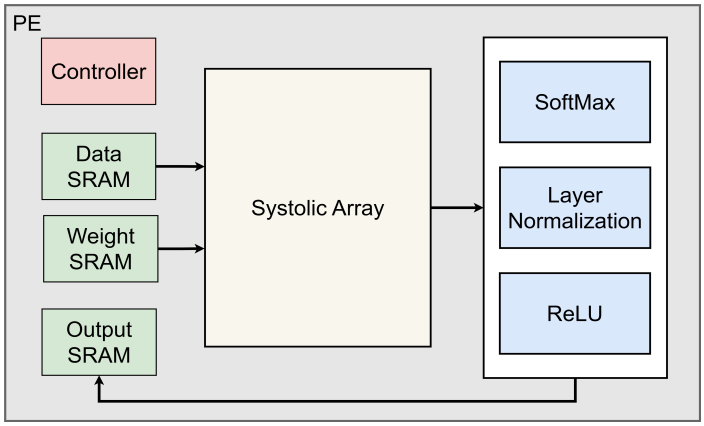

Fig. 3. The architecture of a specialized Processing Element (PE) with its internal controller, SRAMs, systolic array, and nonlinear units.

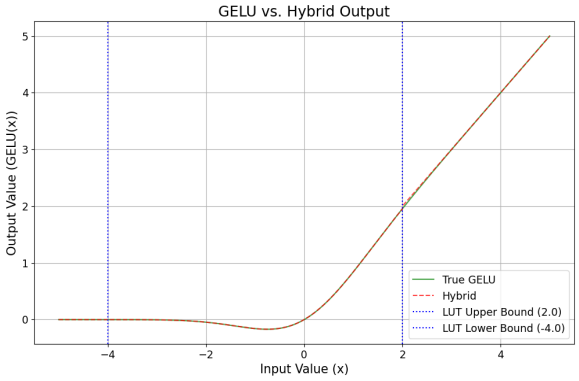

Fig. 4. The Hybrid GELU method accurately approximates the true GELU curve in its critical -4.0 to 2.0 range, while reverting to ReLU behavior elsewhere.

are placed near on-chip memories to reduce transmission congestion. The remaining PEs contain only systolic arrays to conserve area. This design, combining a flexible interconnect with a specialized PE layout, enables NOCTUA to dynamically reconfigure dataflows according to operational needs. For instance, during a large matrix multiplication, the controller can perform data tiling in advance, while the NoC can use efficient multicast to distribute shared blocks to PE and then effectively aggregate the partial sums. This adaptive data path optimization significantly reduces unnecessary data detours and execution latency, ensuring high operational efficiency under varied computational loads.

B. Hardware-Friendly Nonlinear Function Implementation

NOCTUA accelerates key nonlinear Transformer functions, which are often computationally expensive in hardware. It employs dedicated, hardware-friendly units for Layer Normalization, GELU, and Softmax to improve efficiency without a substantial loss in model accuracy. For the GELU activation function, NOCTUA implements a hardware-friendly variant using an asymmetric hybrid lookup table (LUT) method. While the input value is in the range between -4 to 2, the output value is determined based on the LUT. Otherwise, the ReLU function is employed if the input values are outside this range. Fig. 4 shows that the proposed hybrid GELU-ReLU approach closely approximates the original GELU. In this way, we can significantly reduce the LUT size. In other words, it only needs to store values for the most sensitive input region while minimally impacting overall model accuracy. This method substantially reduces the hardware resources and computation cycles typically associated with the standard GELU function.

The Softmax function is critical for computing attention weights. To achieve a hardware-friendly integer-only implementation, similar to the ITA accelerator [11], NOCTUA employs a novel approach that builds on principles of integer-based Softmax to approximate the operation using scaled arithmetic and bit-shifts. This method can be formulated to

$$\begin{aligned}\text{softmax}(x) = \Sigma_{\text{inverse}} \gg \\ ((\max(x_q) - x_{qi}) \gg (B - \log_2 B - \log_2 \alpha)).\end{aligned} \tag{1}$$

For each row of attention scores, the dynamic range is calculated and used to select an adaptive factor, denoted as α, from a predefined set using bucketization. This factor modulates the overall scaling factor with the formula $\epsilon = \alpha \cdot \frac{B}{2^B}$, tailoring the Softmax calculation to the specific distribution of scores. Dynamic scaling is crucial for maintaining differentiability among scores, especially in no-padding scenarios where valid scores may be numerically close. The process derives an integer shift amount for each score by comparing its difference from the maximum score to the dynamic scaling factor ϵ, followed by efficient bit-shift operations to approximate the relative weight of each score. These approximated weights are accumulated to form a denominator, with its scaled integer inverse calculated to generate integerized output probabilities. This method avoids complex operations like exponentiation or logarithms, relying on integer arithmetic, bit shifts, and simple comparisons, significantly reducing hardware complexity, area, and latency while effectively handling various input distributions.

The Layer Normalization is also a key component in Transformers for stabilizing training and enhancing performance. and the formula is defined to

$$y = \frac{x - \mu}{\sqrt{\sigma^2 + \delta}} \cdot \gamma + \beta. \tag{2}$$

Note that the μ and σ^2 are the mean and variance of the input x, respectively. Because of the involved operations (e.g., mean, variance, square root, and division), it is challenging to realize. The primary computational challenge lies in the $1/\sqrt{\sigma^2 + \delta}$ term. To mitigate the design challenge, we address this by employing a Lookup Table (LUT) assisted iterative inverse square root method, evaluated using bfloat16 precision in simulation. This method first utilizes a Piecewise Linear (PWL) approximation, with coefficients pre-stored in a small LUT, to obtain an initial estimate s_0 for $1/\sqrt{V}$. Subsequently, this estimate is refined through a small number of Newton-Raphson iterations, typically 1 to 2, performed at bfloat16 precision. The iteration formula is

$$s_{k+1} = 0.5 \cdot s_k \cdot (3 - V \cdot s_k^2). \tag{3}$$

Finally, the centered input is multiplied by the refined s_k and the learnable gain parameter γ, followed by the addition of the bias parameter β. This method replaces complex square root and division operations with LUT lookups and iterative multiplications and additions, thereby significantly reducing computational complexity and being more amenable to efficient hardware implementation.

IV. Experimental Results and Analysis

NOCTUA is designed for implementation using TSMC 40nm CMOS technology, featuring a 4x4 heterogeneous Processing Element (PE) array interconnected by a network-on-

TABLE I
HARDWARE COMPARISON WITH STATE-OF-THE-ART TRANSFORMER ACCELERATORS

Feature	ITA [11]	RAWAtten [12]	DTQAtten [14]	NOCTUA (Ours)
Technology [nm]	22	40	40	40
Area [mm^2]	0.407	1.64	1.41	45.92
Frequency [MHz]	500	1000	1000	1000
Data formats	INT8	INT8	INT8/INT4	INT8/BF16
Number of MAC units	1024	1024	3168	16384
On-chip memory [KB]	67.8	213	-	1024
Power [mW]	121	660	733.8	31152.3
Throughput [TOPS]	1.02	0.77	0.95	32.77
Energy efficiency [TOPS/W]	8.46	1.17	1.3	1.05
Area efficiency [TOPS/mm^2]	2.52	0.47	0.68	0.71

chip (NoC) and managed by a central controller. The accelerator has an area of 45.92 mm² and consumes around 31.15 W. We evaluated the fidelity of the hardware-friendly nonlinear functions within NOCTUA against their FP32 software counterparts. The asymmetric hybrid LUT-based GELU approximation achieved a Mean Absolute Error (MAE) of 1.6e-2. The dynamic range-adaptive integer Softmax tailors its scaling factor to the input score distribution, effectively maintaining attention fidelity. Additionally, the Piecewise Linear (PWL) assisted Newton-Raphson Layer Normalization, simulated with bfloat16 precision, demonstrated a MAE of 2.8e-3 and a MaxAE of 7.8e-3. Notably, we evaluated NOCTUA's inference performance on classic NLP tasks. Benchmarked with the BERT-Base model, our design achieves 92.4% accuracy on the IMDb sentiment analysis task and 86.6% accuracy on the GLUE MRPC semantic matching task. Compared to the FP32 software baseline, the accuracy degradation is between 0.6% and 1.1%. This strong balance between model accuracy and hardware efficiency confirms that NOCTUA can effectively handle diverse tasks for real-time applications.

Table I presents a comparison of NOCTUA with other state-of-the-art Transformer accelerators. While NOCTUA's power consumption is higher, this reflects a design trade-off to achieve a significantly greater throughput and architectural flexibility. Specifically, this power budget supports a higher performance than the benchmarks, making it suitable for high-demand applications. Although NOCTUA's energy efficiency of 1.05 TOPS/W and area efficiency of 0.71 TOPS/mm^2 are comparable to other accelerators, its primary advantage is its reconfigurable NoC architecture and heterogeneous PE design. This adaptability, combined with high throughput, makes it a powerful and effective solution for accelerating the diverse and evolving computational requirements of LLMs.

V. CONCLUSION

This paper introduces NOCTUA, a high-efficiency, reconfigurable Network-on-Chip (NoC) universal accelerator designed to efficiently infer various Transformer models. Its architecture includes a flexible NoC that supports reconfigurable data flow, controller-managed data tiling with multicast capabilities, and heterogeneous Processing Elements (PEs) equipped with specialized nonlinear units. These features significantly enhance PE utilization and minimize data movement. Evaluated using TSMC 40nm technology, NOCTUA achieves a competitive energy efficiency of 1.05 TOPS/W and an area efficiency of 0.71 TOPS/mm², all while maintaining high model accuracy. This demonstrates its potential for accelerating a wide range of Transformer workloads.

REFERENCES

[1] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, pp. 5998-6008.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186.

[3] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), Vienna, Austria (virtual), May 2021.

[4] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient Transformers: A Survey," ACM Comput. Surveys., vol. 55, no. 6, Art. no. 109, Jun. 2023.

[5] T. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901.

[6] V. Sze, Y. -H. Chen, T. -J. Yang and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," in Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, Dec. 2017.

[7] R. Pope et al., "Efficiently Scaling Transformer Inference," in Proceedings of Machine Learning and Systems, vol. 5, pp. 606-624, 2023.

[8] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), Toronto, ON, Canada, 2017, pp. 1-12.

[9] A.Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, Survey of machine learning accelerators, in Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC), Waltham, MA, USA, Sep. 2020, pp. 112.

[10] L. Benini and G. De Micheli, "Networks on chips: A new SoC paradigm," Computer, vol. 35, no. 1, pp. 70–78, Jan. 2002.

[11] G. Islamoglu et al., "ITA: An Energy-Efficient Attention and Softmax Accelerator for Quantized Transformers," 2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Vienna, Austria, 2023, pp. 1-6.

[12] W. Li, Y. Luo, and S. Yu, "RAWAtten: Reconfigurable Accelerator for Window Attention in Hierarchical Vision Transformers," in 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2023, pp. 1-6.

[13] T. Li, F. Zhang, X. Fan, J. Shen, W. Guo, and W. Cao, "Unified Accelerator for Attention and Convolution in Inference Based on FPGA," in 2023 IEEE International Symposium on Circuits and Systems (ISCAS), 2023, pp. 1-5.

[14] T. Yang et al., "DTQAtten: Leveraging Dynamic Token-based Quantization for Efficient Attention Architecture," in Design, Automation & Test in Europe Conference & Exhibition (DATE), Antwerp, Belgium, 2022, pp. 700-705.