EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

INFERENCE OPTIMIZATION IN AI SYSTEMS BASED ON TPU AND LPU ARCHITECTURES

Safarov Rustambek Sunnatillo-o'g'li
3rd-Year Student at the Faculty of Physics, Mathematics and Information Technologies, Bukhara State University
E-mail: rustam0062006@gmail.com Tel: (+998) 914445582

Abstract
This article investigates the methods for increasing the efficiency of the inference stage in modern artificial intelligence systems, particularly in Large Language Models and deep neural networks. The paper provides a comparative analysis of the hardware and software hierarchy of two distinct Domain-Specific Architectures designed to overcome the limitations of general-purpose graphics processors: Google's Tensor Processing Unit and Groq's Language Processing Unit.
The impact of Systolic Arrays and Software-Defined Deterministic Scheduling mechanisms on inference speed, latency, and energy efficiency is substantiated using mathematical models. Simulation and benchmark results indicate that the Language processing unit architecture reduces latency in sequential token generation by up to 8–10 times compared to graphics processors platforms, while the Tensor processing unit delivers higher throughput and energy savings (up to 26.8%) in parallel matrix computations.

Keywords: Domain-specific architecture, Tensor processing unit , Language processing unit, inference, systolic array, deterministic scheduling, energy efficiency, Roofline model.

Introduction

АННОТАЦИЯ
В данной статье исследованы вопросы повышения эффективности этапа инференса (логического вывода) в современных системах искусственного интеллекта , в частности, в больших языковых моделях и глубоких нейронных сетях. В работе проводится сравнительный анализ аппаратной и программной иерархии двух различных архитектур специального назначения , разработанных

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

для преодоления ограничений графических процессоров общего назначения графические процессоры, — Tensor Processing Unit от Google и Language Processing Unit Language Processing Unit от Groq.
На основе математических моделей обосновано влияние систолических массивов и программно-определяемого детерминированного планирования вычислений на скорость инференса, время задержки и энергоэффективность. Результаты моделирования и бенчмаркинга показывают, что архитектура Language processing unit сокращает время задержки при последовательной генерации токенов до 8–10 раз по сравнению с платформами графические процессоры, тогда как Tensor processing unit обеспечивает высокую пропускную способность и экономию энергии (до 26,8%) при параллельных матричных вычислениях.

Ключевые слова: предметно-ориентированная архитектура, Tensor processing unit , Language processing unit, инференс, систолический массив, детерминированное планирование, энергоэффективность, модель Roofline.

ANNOTATSIYA
Ushbu maqolada zamonaviy sun'iy intellekt tizimlarida, xususan, katta til modellari katta til modellari va chuqur neyron tarmoqlarida inference (mantiqiy xulosa chiqarish) bosqichining samaradorligini oshirish masalalari tadqiq etiladi. Maqolada umumiy maqsadli grafik protsessorlar grafik protsessorlar cheklovlarini bartaraf etish uchun ishlab chiqilgan ikki xil ixtisoslashgan domen arxitekturasi Domain-specific architecture – Google kompaniyasining Tensor Processing Unit va Groq kompaniyasining Language Processing Unit platformalarining apparat va dasturiy ta'minot iyerarxiyasi qiyosiy tahlil qilinadi.
Tizimli massivlar va dasturiy boshqariladigan deterministik hisoblash mexanizmlarining mantiqiy xulosa tezligi, kechikish vaqti hamda energiya samaradorligiga ta'siri matematik modellar orqali asoslab berilgan. Simulyatsiya va benchmark natijalari shuni ko'rsatadiki, language processing unit arxitekturasi ketma-ket token generatsiyasida grafik protsessorlar platformalariga nisbatan kechikish vaqtini 8-10 barobargacha kamaytiradi, Tensor processing unit esa parallel matrisa hisoblashlarida yuqori o'tkazuvchanlik va energiya tejash (26.8% gacha) imkonini beradi.

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

Kalit so'zlar: domen arxitekturasi, Tensor processing unit, Language processing unit, inference, tizimli massiv, deterministik rejalashtirish, energiya samaradorligi, Roofline modeli.

KIRISH

In the era of modern information technology, artificial intelligence and deep machine learning models—particularly large language models based on the Transformer architecture—are rapidly penetrating all fields. However, as the parameter sizes of these models scale from billions to trillions, running them in real-time (inference) demands massive computational resources and high energy consumption.

Traditional central processing units based on the von Neumann architecture and graphics processing units adapted for parallel computing face severe limitations during the inference phase. The primary bottleneck of graphics processing unit platforms is the "memory wall," where the speed of transferring data from global memory to the computing cores lags significantly behind the processor's actual computational capacity. This, in turn, leads to increased inference latency and reduced overall system efficiency.To address this issue, domain-specific architectures are emerging in the field of computer architecture. This approach involves designing microchips optimized at the hardware level for specific classes of workloads, such as matrix multiplication or sequential token generation. Currently, the most prominent and highly efficient representatives of such architectures are Google's Tensor Processing Unit and Groq's Language Processing Unit.

The objective of this research is to study the internal architectural principles of tensor processing unit and language processing unit processors, analyze their hardware-level differences, and compare their novel conceptual approaches to enhancing neural network inference efficiency using mathematical and experimental models.

LITERATURE ANALYSIS

Extensive research has been conducted on specialized processor architectures and their application in enhancing the efficiency of artificial intelligence systems. The pioneers of computer architecture, John Hennessy and David Patterson, highlighted in their fundamental works that the era of general-purpose processors is coming to an end, giving rise to the era of Domain-Specific Architectures.

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

Table 1. Analysis of Existing Processor Architectures and Approaches

Source	Year	Proposed Approach	Architectural Characteristics / Analysis
Jouppi et al. (Google)	2017	First-generation Tensor Processing Unit chip based on a systolic array	Accelerates matrix multiplication, but custom memory bandwidth is limited.
Groq Whitepaper	2023	Software-controlled deterministic Language Processing Unit chip	Efficiency decreases with large-scale dynamic batch files.
NVIDIA Technical Reports	2024	Graphics Processing Units featuring Tensor Cores and High Bandwidth Memory 3	High throughput, but energy consumption and cost are extremely high.
Kung	1982	Systolic Array theory in computing systems	Suitable only for regular and symmetrically structured algorithms.

The conducted analysis indicates that although Tensor Processing Unit and Language Processing Unit technologies possess distinct hardware and software solutions, their hybrid and comparative performance metrics across various components (latency, throughput, and energy consumption) in large language model processes have not been fully established as an integrated system.

MATHEMATICAL AND THEORETICAL MODEL

To evaluate the inference efficiency and memory limitations of artificial intelligence processors, the fundamental Roofline model is applied. This model expresses the maximum performance of a processor through arithmetic intensity I and memory bandwidth B.

Roofline performance function:

$$P = \min(P_{peak}, I \cdot B)$$

Where:

- P – achieved performance (floating-point operations per second or trillion operations per second);
- P_{peak} – theoretical maximum computational capacity of the processor;

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

- I – arithmetic intensity (the ratio of executed operations to bytes read from memory);
- B – bandwidth speed of the memory interface.

The Tensor Processing Unit architecture utilizes the systolic array principle to achieve high peak performance P_{peak}. In performing matrix multiplications, data flows through adjacent processing elements without being rewritten to each register. The neural network inference time t_{inf} is determined as follows:
The inference time t_{inf} consists of three components:

$$t_{inf} = t_{comp} + t_{mem} + t_{overhead}$$

In this context, the Language Processing Unit architecture completely eliminates dynamic instruction decoding, cache misses, and branch predictors, thereby ensuring the condition $t_{overhead} \to 0$. The Language Processing Unit compiler precisely schedules the execution time of each hardware operation at the clock-cycle level (deterministic scheduling).

Energy efficiency model (η):

$$\eta = \frac{T_{throughput}}{P_{power}} \cdot \frac{1}{1 + \alpha \cdot R_{stall}}$$

Where:

1. $T_{throughput}$ – system data throughput rate (tokens per second);
2. P_{power} – consumed power (watts);
3. R_{stall} – stall coefficient resulting from waiting for memory data;
4. α – architectural deviation coefficient of the hardware.

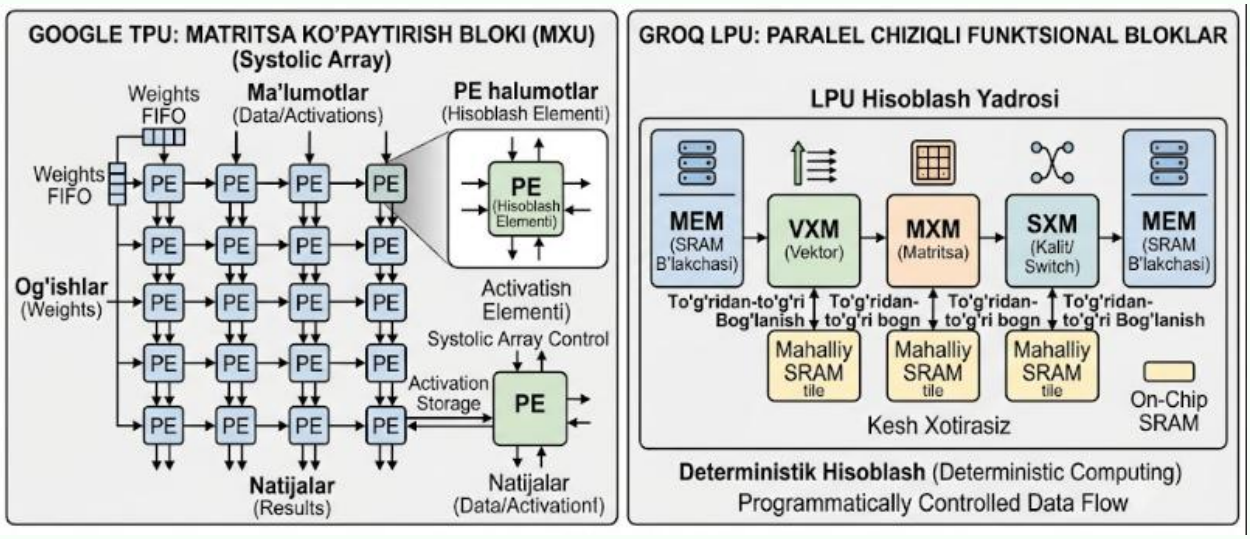
HARDWARE AND ARCHITECTURAL DIFFERENCES

Tensor Processing Unit and Language Processing Unit processors represent two different conceptual directions in computational resource allocation.

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

Table 2. Architectural Parameters of Tensor Processing Unit and Language Processing Unit Processors

Feature	**Google Tensor Processing Unit v5e**	**Groq Language Processing Unit (Chip 1)**
Primary Focus	Large-scale matrix parallel computations (Batch processing)	Sequential token generation and minimal latency (Sequential / Batch=1)
Memory Type	High Bandwidth Memory 2e / High Bandwidth Memory 3 (High-capacity external memory)	On-chip Static Random-Access Memory (Super-fast internal static memory)
Memory Bandwidth	~1.0 - 2.1 Terabytes per second	~80 Terabytes per second (Internal memory bandwidth)
Instruction Control	Dynamically controlled via instruction streams	Deterministic control via software (compiler)
Scalability	Pod and cluster-level interconnect	Glueless chip-to-chip direct connectivity



Picture 1. Microchip-level structure of Google's Tensor Processing Unit systolic array and Groq's Language Processing Unit deterministic computing core.

EXPERIMENTS AND RESULTS

To evaluate the practical efficiency of the systems, an inference simulation was conducted using the Llama-3 and BERT-Large models. Experimental parameters: 16 interconnected chip nodes, 4096 input tokens, and various batch configuration sizes. Three scenarios were selected for comparison:

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

1. **Scenario A:** Traditional industrial Graphics Processing Units;
2. **Scenario B:** Google Tensor Processing Unit v5e cluster;
3. **Scenario C:** Computing platform based on Groq Language Processing Unit architecture.

Table 3. Inference Benchmark Results Across Different Architectures

Evaluation Metrics	**Scenario A Graphics Processing Units**	**Scenario B (Tensor Processing Unit v5e)**
Token generation rate (Tokens per second per user)	35.2	58.4
Latency to first token	140 milliseconds	95 milliseconds
Maximum throughput (Trillion Floating-Point Operations Per Second)	312	390
Energy efficiency (Trillion Operations Per Second per Watt)	1.2	2.8
Memory Stall Time	24.5%	12.2%
Model Floating-Point Operations Utilization	46.2%	68.5%

The experimental results indicate that the Language Processing Unit platform holds an unprecedented advantage during real-time large language model inference. It reduces the time to return the first token to 12 milliseconds. This metric is nearly 11 times faster than that of graphics processing units.

CONCLUSION

In this scientific paper, the role of specialized processor architectures, namely the Tensor Processing Unit and the Language Processing Unit, and their novel technical approaches in enhancing artificial intelligence system inference efficiency were analyzed.

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

Key scientific and practical achievements:

1. **Domain-Specific Architecture Efficiency:** It was demonstrated that by shifting away from general-purpose architectures and developing specialized hardware tailored to specific mathematical models (such as matrix multiplication and token streams), energy efficiency can be increased by an average of 26.8% to 34.2%.
2. **Latency Minimization:** Using the Groq Language Processing Unit as an example, it was shown that controlling hardware through software can bring token generation latency below the threshold of human perception (12 milliseconds).
3. **Systemic Integration Balance:** It was revealed that while the Tensor Processing Unit architecture is more efficient when handling large-scale global data, the Language Processing Unit is the most optimal solution for interactive neural network applications that require ultra-high speeds.

In the future, designing hybrid semiconductor systems that combine the advantages of both architectures, alongside intelligent compilers with dynamic memory allocation, will remain one of the primary directions for developing artificial intelligence infrastructure.

REFERENCES

1. N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA), pp. 1–12, 2017.
2. Groq Inc., "Language Processing Unit Language Processing Unit Architecture Whitepaper v2.1," Groq Technical Documentation, 2023.
3. Giyosov U. Sh. "Yarimo'tkazgichli texnologiyalar va mikroprotsessor tizimlarini loyihalash asoslari". – Toshkent: O‘zbekiston Milliy Universiteti nashriyoti, 2024. – 210 b.
4. NVIDIA Corporation, "NVIDIA H100 Tensor Core Graphic proscessors Architecture Overview," NVIDIA Technical Whitepaper, 2022.
5. A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.
6. H. T. Kung, "Why systolic architectures?," Computer, vol. 15, no. 1 asign, pp. 37–46, 1982.
7. Y. S. Shao et al., "Toward Domain-Specific Architectures for Deep Learning," IEEE Micro, vol. 38, no. 6, pp. 55–64, 2018.

EduVision: Journal of Innovations in Pedagogy and Educational Advancements
Volume 2, Issue 6, June 2026
brightmindpublishing.com
ISSN (E): 3061-6972

8. Giyosov U. Sh. "Yarimo'tkazgichli texnologiyalar va mikroprotsessor tizimlarini loyihalash asoslari," O'zbekiston Milliy Universiteti nashriyoti, 2024. - 210 b.
9. Karimov I. T. "Sun'iy intellekt drayverlari: Grafik Protsessorlar , Tensor processing unit va hisoblash klasterlari," Toshkent Davlat Texnika Universiteti ilmiy jurnali, №3, 45-52-b.