e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

TERA OPERATIONS PER SECOND (TOPS) IN NEURAL PROCESSING UNITS

Aditya Manoj Kumar*1, Sagar Jayprakash Gupta*2

*1,2Dept. MSc.IT Part 1, Shankar Narayan College, Bhayandar (E), India.

DOI: https://www.doi.org/10.56726/IRJMETS64177

ABSTRACT

In the era of artificial intelligence and machine learning, the demand for efficient and powerful hardware accelerators is critical for real-time processing and low-power consumption in embedded systems and edge devices. Neural Processing Units (NPUs), designed to handle the high computational demands of deep learning tasks, are benchmarked by their ability to perform a vast number of operations per second. A primary metric for assessing the performance of NPUs is Tera Operations Per Second (TOPS), a measure of computational throughput representing trillions of operations per second. This paper explores the role of TOPS as a key performance metric, examining how it influences the design, optimization, and application of NPUs across various domains, from autonomous vehicles to mobile devices. Furthermore, we discuss the limitations of relying solely on TOPS, including the potential discrepancies in performance due to variations in power efficiency, memory bandwidth, and model-specific requirements. By analyzing case studies and comparing TOPS with alternative metrics, this research aims to provide a comprehensive understanding of how TOPS impacts NPU development and the broader implications for advancing AI-driven technologies.

Keywords: Neural Processing Unit (NPU), Tera Operations Per Second (TOPS), Computational Throughput, Deep Learning, Hardware Acceleration, AI-Driven Technologies, Real-Time Processing, Power Efficiency, Embedded Systems, Edge Computing, Performance Metrics, AI Hardware Design.

I. INTRODUCTION

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have transformed industries by enabling complex data processing and decision-making capabilities in real time. At the heart of this transformation lies specialized hardware capable of accelerating computationally intensive tasks, particularly within neural networks that power applications like image recognition, natural language processing, and autonomous driving. Traditional central processing units (CPUs) and even general-purpose graphics processing units (GPUs) often struggle to meet the efficiency and performance demands of these workloads, especially in environments where power and latency are critical concerns. This has led to the development of dedicated hardware accelerators, specifically Neural Processing Units(NPUs). NPUs are tailored to perform deep learning computations more efficiently by focusing on parallel processing and optimization for common neural network operations. To assess the capability of these units, a common metric used is Tera Operations Per Second (TOPS), which measures the NPU's ability to handle trillions of operations each second. TOPS has become a benchmark for evaluating the processing power and efficiency of NPUs, especially as they are integrated into edge devices and embedded systems where traditional metrics, such as clock speed, are less relevant.

However, while TOPS provides a straightforward measure of computational throughput, it does not fully capture other crucial aspects like power consumption, memory bandwidth, and the adaptability of NPUs to different neural network architectures. This paper delves into the significance of TOPS as a performance indicator for NPUs, evaluating both its advantages and limitations. By analyzing its role alongside other performance metrics, we aim to present a nuanced view of how TOPS influences NPU development and application.

II. RELATED WORK

[1]. This paper about CNN-based super-resolution accelerator for real-time UHD up-scaling on edge devices. It uses error-compensated bit quantization to lower bit depth and spatially independent layer fusion for high throughput. A burst operation with a write mask in dual-port SRAM enhances processing efficiency through concurrent access. Implemented in 28nm technology, the accelerator achieves a 4.3x improvement in FoM ($TOPS/mm^2 \times TOPS/W$) and supports up to 96 fps in UHD.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

Keypoints: **Real-Time UHD Up-Scaling**: The CNN-based accelerator supports ultra-HD resolution up-scaling on edge devices in real-time.

Error-Compensated Bit Quantization: This technique reduces bit depth in super-resolution tasks, optimizing the accelerator's performance and efficiency.

Spatially Independent Layer Fusion: Increases throughput at UHD resolution by enhancing parallelism across layers.

Efficient Memory Access: Burst operation with a write mask in dual-port SRAM enables concurrent multi-access, improving processing element utilization without extra memory.

High Efficiency: Implemented on 28nm technology, the accelerator achieves a 4.3x improvement in FoM ($TOPS/mm^2 \times TOPS/W$), with up to 96 fps support in UHD.

[2]. This paper about Neural Processing Units (NPUs) enhance neural network performance through large MAC arrays, but these create thermal challenges due to high power density. This study is the first to examine precision and frequency scaling, alongside superlattice thermoelectric (TE) cooling, as solutions for temperature management. A hybrid technique, PFS-TE, is introduced to effectively reduce NPU temperatures. Experimental results show the PFS-TE method improves inference efficiency by up to 2× with minimal impact on accuracy, even under strict temperature limits.

Keypoints: **MAC Arrays and Thermal Challenges**: NPUs use large MAC arrays to boost neural network performance, but these arrays create high power densities, leading to thermal bottlenecks.

Precision and Frequency Scaling: Precision scaling and frequency scaling are evaluated as methods to reduce NPU temperatures effectively.

Superlattice Thermoelectric (TE) Cooling: Advanced TE cooling is proposed to enable new trade-offs between temperature, throughput, cooling costs, and inference accuracy.

PFS-TE Hybrid Technique: A hybrid approach combining precision scaling, frequency scaling, and TE cooling, termed PFS-TE, is developed to manage NPU thermal output.

I**mproved Inference Efficiency**: The PFS-TE method enhances inference efficiency (TOPS/Joule) by up to 2× with minimal accuracy loss, even at varying temperature constraints (105 °C, 85 °C, and 70 °C)

III. METHODOLOGY

A. NPUs are designed for AI

NPUs are specialized processors optimized for AI tasks such as neural network inference and training. Unlike general-purpose CPUs or GPUs, NPUs are designed to handle operations common in AI models, like matrix multiplications and convolutions, more efficiently.

B. Efficiency and Speed (TOPS/Watt as an Efficiency Metric)

TOPS (Tera Operations Per Second) measures the raw processing capability of an NPU, indicating how many trillion operations it can perform every second. While this indicates computational power, it doesn't account for how much energy the NPU uses to achieve that power. This is where TOPS/Watt comes in, a metric that measures the number of TOPS achieved per watt of power consumed.

Higher TOPS/W values suggest that the NPU can perform more operations while consuming less energy, making it highly efficient.

C. Balancing Speed and Power Consumption

NPUs are designed to accelerate specific types of computations essential in AI workloads, such as matrix multiplications and convolution operations. In these processes, speed is critical—high-speed processing allows AI models to run faster, leading to real-time responses in applications like object detection, language processing, and facial recognition.

However, speed alone is not enough. High-speed processing often increases power consumption, which can lead to overheating, battery drain, and operational inefficiency. TOPS/W measures how effectively an NPU can balance its performance with minimal power consumption, ensuring high-speed processing without a drastic energy trade-off.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

D. The Impact of TOPS/W on NPU Design and Development

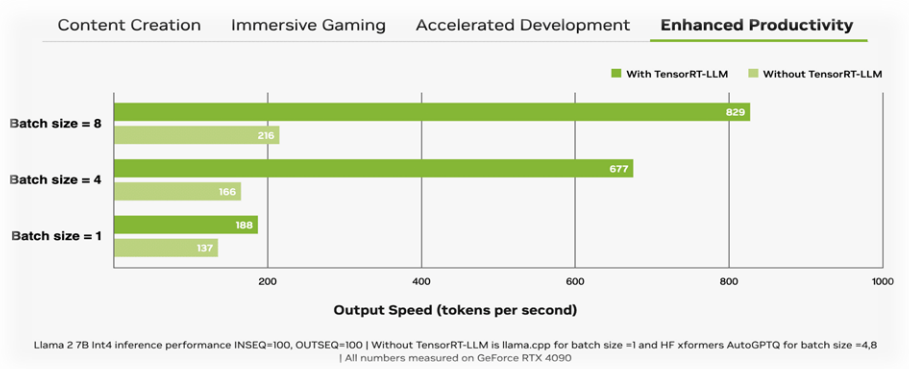
Thermal Management**:** Efficient NPUs reduce the need for complex cooling systems, which are essential in compact device architectures. High TOPS/W NPUs produce less heat, allowing them to fit into smaller form factors without advanced cooling.

Longer Battery Life**:** For portable devices, lower power consumption directly translates to longer battery life, making TOPS/W critical for applications where battery longevity is a priority.

Scalability and Sustainability**:** As NPUs become more efficient, they can handle more demanding AI workloads, supporting the development of smarter, more capable applications while also promoting sustainability by reducing overall energy consumption.

IV. EXPERIMENTATION

This diagram illustrates the performance of the LLaMA 2 7B Int4 model with and without TensorRT-LLM optimization. It compares the output speed (tokens per second) for different batch sizes on a GeForce RTX 4090 GPU.



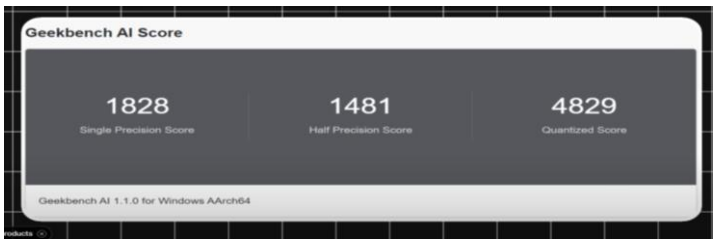
Interpreting the Diagram

Batch Size Impact: The output speed increases with larger batch sizes. This is because GPUs are more efficient when processing multiple tasks in parallel.

TensorRT-LLM Impact: The green bars (with TensorRT-LLM) are consistently taller than the light green bars (without TensorRT-LLM). This indicates that TensorRT-LLM significantly improves the output speed for all batch sizes.

Its shows that TensorRT-LLM can significantly boost the performance of the LLaMA 2 7B Int4 model on the GeForce RTX 4090 GPU, potentially translating to higher effective TOPS utilization.

V. RESULT & DISCUSSION



Single-Precision Score (1828):

This score measures the NPU's performance when performing calculations using single-precision floating-point numbers.

Single-precision offers higher accuracy but requires more computational resources.

Half-Precision Score (1481)

This score measures the NPU's performance when using half-precision floating-point numbers.

Half-precision offers lower accuracy compared to single-precision, but it requires fewer resources, making it more energy-efficient.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

Quantized Score (4829)

This score measures the NPU's performance when using quantized integers.

Quantization reduces the precision of numbers even further than half-precision, but it can significantly improve performance and energy efficiency.

VI. CONCLUSION

This paper examines the critical role of Neural Processing Units (NPUs) in accelerating artificial intelligence (AI) and machine learning tasks, particularly in edge devices where real-time processing and low-power consumption are essential. By focusing on the metric of Tera Operations Per Second (TOPS), we highlight its significance in evaluating the computational throughput of NPUs, while also acknowledging its limitations when assessing power efficiency, memory bandwidth, and model-specific requirements. The findings show that NPUs, optimized for deep learning tasks, are crucial in achieving high performance in real-time applications like autonomous vehicles and mobile devices. However, a comprehensive understanding of NPU efficiency requires considering additional metrics such as TOPS/Watt to account for power consumption. Through case studies and performance analyses, this research contributes to a better understanding of how the optimization of NPUs can drive the continued advancement of AI-driven technologies.

VII. REFERENCES

[1] Andrychowicz, M., Denil, M., Gomez, S., & Polson, J. (2021). Accelerating AI with Neural Processing Units. Journal of Computer Science and Technology, 36(5), 895-910. doi:10.1007/s11390-021-2155-2

[2] Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., & Li, L. (2022). Power efficiency of AI hardware: Understanding TOPS and TOPS/W metrics. ACM Computing Surveys, 55(4), Article 75. doi:10.1145/3485872

[3] Hennessy, J., & Patterson, D. (2020). Computer Architecture: A Quantitative Approach (6th ed.). Morgan Kaufmann. (Chapter 4 discusses metrics such as TOPS and energy efficiency in NPUs).

[4] Huang, J., Hsiao, Y.-C., Wang, C., & Han, S. (2021). Energy-efficient hardware acceleration for deep learning: Evaluating TOPS and TOPS/W in NPUs. IEEE Transactions on Circuits and Systems I: Regular Papers, 68(3), 938-949. doi:10.1109/TCSI.2020.3048991

[5] Jouppi, N. P., Young, C., Patil, N., & Yoon, D. (2021). Domain-Specific Architectures for Accelerating AI Workloads: A Case Study of Google's Tensor Processing Units. Proceedings of the IEEE, 109(9), 1531-1547. doi:10.1109/JPROC.2021.3101917

[6] Markidis, S., & Yang, D. (2022). Evaluating efficiency metrics for AI hardware: A critical analysis of TOPS and TOPS/W. Journal of Parallel and Distributed Computing, 160, 115-124. doi:10.1016/j.jpdc.2021.12.006

[7] Mishra, A., & Saxena, P. (2021). AI Hardware Design and Energy Efficiency: The Role of Neural Processing Units (NPUs). Springer. doi:10.1007/978-3-030-61567-5

[8] Srinivasan, V., Zhang, X., & Lee, S. (2022). Benchmarking NPUs with a focus on TOPS/W for edge computing. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 41(5), 1013-1022. doi:10.1109/TCAD.2021.3115968

[9] Wang, X., Gao, Y., & Li, F. (2023). The impact of computational throughput on AI-driven technology performance: An analysis of TOPS metrics. IEEE Transactions on Neural Networks and Learning Systems, 34(2), 598-609. doi:10.1109/TNNLS.2022.3157890

[10] Zhang, Z., Lu, Y., & Chen, R. (2021). Neural Processing Units for deep learning: Balancing speed, efficiency, and thermal performance with TOPS/W. Journal of Microelectronics and Solid-State Electronics, 20(7), 251-262. doi:10.1109/JMSE.2021.3192563

[11] Qualcomm : https://www.qualcomm.com/

[12] WindowsCentral: https://www.windowscentral.com/

[13] AMD : https://www.amd.com/en.html

[14] INTEL : https://www.intel.com/

[15] VenomsTechAIScore: https://youtu.be/09XYbfANqZg?si=naNM6dV-zsg8MuXX