Scalability of the ARM Nvidia Grace Superchip for Deep Learning Applications [*]

Thiago Araújo[1], **Philippe Navaux**[1]

[1]Informatic Institute – Federal University of Rio Grande do Sul (UFRGS)

Abstract. *The advance of Artificial Intelligence (AI) has heightened the demand for computational resources, presenting challenges in scalability and energy efficiency. The low-power ARM architecture is a promising option for AI workloads. This study assesses the scalability of the ARM Nvidia Grace Superchip in Deep Learning (DL) using a model that predicts referrals for severe diabetic retinopathy. We analyze performance in terms of speedup and energy consumption across 1 to 144 cores. Although speedup increases with more threads, the gains are less pronounced at higher core counts due to resource saturation. Energy consumption, however, significantly decreases, showing a 95% reduction when utilizing all cores. These results emphasize the Superchip's potential for scalable and energy-efficient AI, especially in demanding applications. Future research will focus on memory usage, latency, and distributed learning.*

1. Introduction

The quick growth of data and technological advancements have resulted in the widespread adoption of AI across various fields. However, training and operating AI models demand substantial computing capability, which leads to high energy consumption. As the use of AI applications continues to expand, scaling these models has become a significant challenge. Meanwhile, the increasing emphasis on sustainability has prompted researchers to develop AI models that scale efficiently and minimize their environmental impact through optimized architectures and energy-efficient computing [Dash 2025].

The ARM architecture is recognized for its low-power design, making it particularly suitable for AI applications. The ARM Nvidia Grace Superchip is specifically engineered for AI workloads. Although several studies have investigated the scalability of this architecture, we have not found any research that focuses on its scalability in the context of AI applications [Banchelli et al. 2024] [Ruhela et al. 2024].

This work analyzes the scalability of the ARM architecture for deep learning applications, focusing on a model that predicts the referral of patients with severe diabetic retinopathy to a specialist. Section II details the methodology and presents the results, while Section III discusses the conclusion and outlines directions for future work.

2. Results

The scalability evaluation for ARM architecture involves training and validating the MobileNet algorithm across different thread counts, starting from the maximum of 144 cores (two sockets with 72 cores each) and halving the count successively to 72, 36, 18, 9, and 1 thread. The experiments were executed in only one Nvidia Grace Superchip.

[*]This work was supported by CAPES - Funding Code 001, CNPq/MCTI/FNDCT - Universal 18/2021 under number 406182/2021-3, CIARS RITEs/FAPERGS and CI-IA FAPESP-MCTIC-CGI-BR projects.

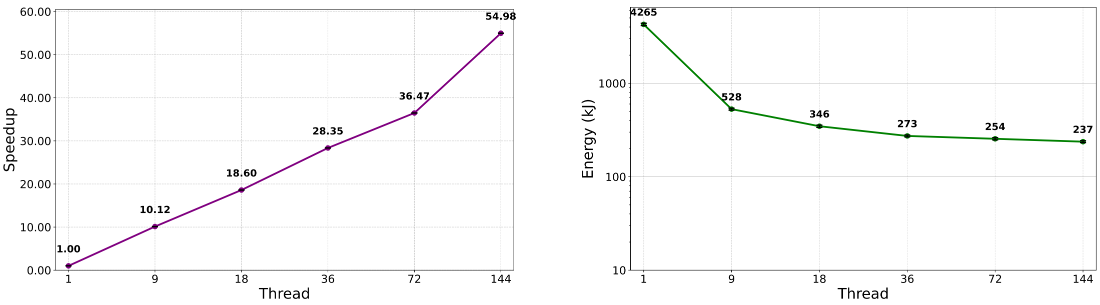

Figure 1. Speedup

Figure 2. Energy Consumption

Figure 3. Scalability of ARM Nvidia Grace in DR application

The first metric analyzed for scalability is speedup, which is calculated by dividing the training and validation time by the longest recorded time. Figure 1 shows the speedup by thread number. The most significant increase in speedup occurred when scaling from 1 to 9 threads due to better workload distribution. The speedup gains continued but at a decreasing rate, with a final speedup of 54.98 using all 144 cores. Figure 2 allows scalability to be analyzed from an energy perspective for ARM. The most substantial reduction in energy consumption occurred when scaling from 1 to 9 threads, following the trend observed in speedup. Beyond this point, further reductions were more modest. Using all 144 cores resulted in an energy consumption of 237 kJ, representing a 95% reduction compared to the energy consumed with a single core.

3. Conclusion

The study examines the model's scalability on the ARM Nvidia Grace Superchip architecture, evaluating speedup and energy consumption over different thread counts. Speedup is proportionally lower when using fewer threads due to higher resource utilization. This trend is also reflected in energy consumption, with the most significant reductions occurring at lower thread counts. As the number of threads approaches the maximum available, resource utilization decreases, leading to a more minor increase in speedup and a more pronounced reduction in energy consumption. Future work should focus on optimizing ARM scalability by analyzing latency and memory usage when handling larger datasets. In addition, scalability with more nodes should be evaluated using distributed learning.

References

Banchelli, F., Vinyals-Ylla-Catala, J., Pocurull, J., Clascà, M., Peiro, K., Spiga, F., Garcia-Gasulla, M., and Mantovani, F. (2024). Nvidia grace superchip early evaluation for hpc applications. In *Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region Workshops*, pages 45–54.

Dash, S. (2025). Green ai: Enhancing sustainability and energy efficiency in ai-integrated enterprise systems. *IEEE Access*.

Ruhela, A., Cazes, J., McCalpin, J., Del-Castillo-Negrete, C., Li, J., Liu, H., Chen, H., Lu, C.-Y., Milfeld, K., Zhang, W., et al. (2024). Performance analysis of scientific applications on an nvidia grace system. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 558–566. IEEE.