Late Breaking Results: Leveraging Approximate Computing for Carbon-Aware DNN Accelerators

Aikaterini Maria Panteleaki*, Konstantinos Balaskas†, Georgios Zervakis†, Hussam Amrouch‡, Iraklis Anagnostopoulos*
*Southern Illinois University Carbondale, †University of Patras, ‡Technical University of Munich

***Abstract*—The rapid growth of Machine Learning (ML) has increased demand for DNN hardware accelerators, but their embodied carbon footprint poses significant environmental challenges. This paper leverages approximate computing to design sustainable accelerators by minimizing the Carbon Delay Product (CDP). Using gate-level pruning and precision scaling, we generate area-aware approximate multipliers and optimize the accelerator design with a genetic algorithm. Results demonstrate reduced embodied carbon while meeting performance and accuracy requirements.**

***Index Terms*—Approximate Accelerators, Embodied Carbon Footprint, Sustainable Computing**

I. Introduction

The rapid growth of machine learning (ML) has driven advances in computing, with specialized hardware accelerators enhancing the efficiency of Deep Neural Networks (DNNs). However, this progress comes with a significant environmental cost, as the embodied carbon footprint from the manufacturing of these accelerators remains largely unexplored. Recent studies [1], [2] highlight that embodied carbon now surpasses operational emissions as a dominant factor in the environmental impact of ML systems, particularly in edge-based applications.

Designing DNN hardware accelerators is challenging due to the wide range of possible hardware configurations and mappings. Key decisions, such as determining the number of Processing Elements (PEs) and setting up local and global memory configurations, greatly affect the accelerator's performance. However, previous works [3] have shown that such accelerators are often overdesigned, providing more performance than necessary for edge applications, and significantly increasing their embodied carbon footprint at the same time. Relaxing performance requirements offers a promising solution that allows the development of carbon-aware designs, that better balance performance and sustainability.

Moreover, DNNs are inherently resilient to computational errors in arithmetic operations, making them an ideal candidate for leveraging approximate computing to reduce embodied carbon. By introducing approximate arithmetic units, which require fewer transistors and have a smaller hardware footprint, it is possible to significantly lower the embodied carbon of DNN accelerators. These approximate units not only reduce the area required for computation, but also free up design space for optimizing memory configurations. Despite the potential benefits, no prior work has explored the use of approximate computing as a means to address the embodied carbon emissions of DNN accelerators.

In this work, we investigate how relaxed constraints along with approximate computing can be systematically applied to balance embodied carbon footprint and performance for DNN inference accelerators.

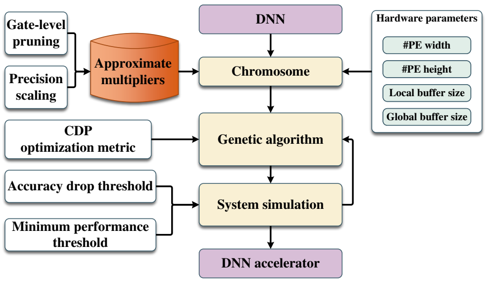


Fig. 1: Overview of the proposed methodology

II. Methodology

The primary objective of our methodology is to design an approximate DNN accelerator and determine its corresponding mapping to optimize the Carbon Delay Product (CDP). CDP is a comprehensive metric that integrates performance and the embodied carbon footprint, offering a holistic assessment of the trade-offs between sustainability and efficiency. By focusing on minimizing the CDP, we aim to design hardware that achieves a balance between performance and carbon emissions, ensuring suitability for edge-based systems.

The embodied carbon estimation of an accelerator, while related to chip area, primarily depends on the die manufacturing process, which incorporates several technology and fab-specific factors beyond area alone. Key contributors include the fabrication facility's attributes, such as its power consumption and the carbon intensity of its electricity grid. Additionally, the technology node used in the fabrication process significantly impacts scaling trends and yield results. For a monolithic DNN accelerator die, the embodied carbon footprint is calculated based on emissions produced during the manufacturing of its logic chip area, using a specified technology node [4]. The total embodied carbon of a chip comprises two main components: the product of the Carbon Footprint Per unit Area (CFPA) of the die and its area (A_{die}), and the product of the CFPA of Silicon (CFPA_{Si}) and the wasted area of the silicon wafer (A_{wasted}) during fabrication, as shown in Eq. 1. The CFPA depends on factors such as the Carbon Intensity of the fabrication facility (CI_{fab}), the Energy consumed per unit Area during manufacturing (EPA), the greenhouse gas emissions (C_{gas}), the carbon impact of raw material procurement (C_{material}), and the yield (Y) of the fabrication process.

$$C_{\text{embodied}} = \text{CFPA} \times A_{\text{die}} + \text{CFPA}_{\text{Si}} \times A_{\text{wasted}} \qquad (1)$$

978-3-9826741-0-0/DATE25/©2025 EDAA

$$\mathrm{CFPA} = \frac{\mathrm{CI_{fab}} \times \mathrm{EPA} + \mathrm{C_{gas}} + \mathrm{C_{material}}}{\mathrm{Y}} \tag{2}$$

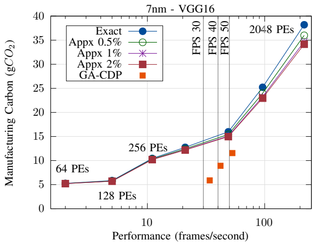
To optimize embodied carbon, while maintaining computational efficiency, we start by generating area-aware approximate multipliers for the MAC units. To achieve this, we apply gate-level pruning and precision scaling approximation techniques to modify the netlist structure or the connections between its gates, effectively reducing the circuit area [5]. These approximations are guided by a multi-objective optimization algorithm that explores the design space to identify near-Pareto-optimal solutions with minimal functional error. The resulting approximate multipliers not only lower the embodied carbon footprint by reducing the required hardware area but also maintain the computational accuracy needed for error-resilient DNN tasks.

In the second step, we integrate these approximate multipliers into exploring hardware configurations and mappings for the accelerator. This involves optimizing key characteristics such as the width and height of the accelerator (number of Processing Elements), local register file sizes, and global buffer capacity. Mapping parameters, including tiling strategies, execution order, and levels of parallelism, are also considered. To navigate this vast design space, we employ a genetic algorithm, with CDP metric as fitness function, to select the Pareto-optimal approximate multipliers from step one and identify the most efficient DNN topology. The optimization process is constrained by thresholds for accuracy drop and performance, measured in inferences per second, ensuring that the design meets the realistic requirements of edge systems. This approach addresses the overdesign issue observed in previous accelerators, resulting in a more sustainable and efficient solution.

III. Evaluation

As a baseline for our design exploration, we use the NVDLA architecture paradigm, which include MAC arrays ranging from 64 to 2048 PEs in powers of 2. The sizes of the local and global convolution buffers scale proportionally with the dimensions of the MAC arrays, as specified by NVIDIA [6]. To evaluate these configurations, we employ two specialized tools: the nn-dataflow [7], to estimate DNN workload performance, and ApproxTrain [8], to calculate the accuracy impact of the approximate multipliers.

Figure 2 illustrates the trade-off between embodied carbon and performance for DNN accelerators running VGG16 at the 7nm technology node. The configurations for the exact (baseline) accelerator show exponential carbon increase as the architecture becomes more compute-intensive. By incorporating approximate units only, while keeping the architecture unchanged (same number of PEs and memory), we achieved a 5% reduction in embodied carbon. In our experiments, we tested approximate units that resulted in accuracy losses of up to 0.5%, 1.0%, and 2.0% (Appx in legend). Similar trends appeared at 14nm and 28nm, as shown in the corresponding table, with gains of up to 12.75%. However, the frames per second (FPS) achieved by large accelerators often far exceed the requirements for edge applications [3]. To address this, we applied realistic performance thresholds of 30, 40, and 50 FPS and utilized our genetic algorithm with approximate multipliers. This approach significantly reduced the embodied carbon footprint, achieving reductions of up to 50%. Each point on the plot (GA-CDP in legend) represents an accelerator design that optimizes the CDP while meeting the specified performance thresholds (30, 40, and 50 FPS).



Carbon Footprint Reduction (%)				
Technology (nm)	Type	Accuracy Drop		
		0.5%	1.0%	2.0%
7	Avg	2.83	4.49	5.17
	Peak	5.78	9.18	10.56
14	Avg	5.58	6.90	8.02
	Peak	8.87	10.98	**12.75**
28	Avg	3.33	5.71	8.44
	Peak	4.60	7.87	11.65

Fig. 2: Embodied Carbon trends for VGG16 across different accuracy and performance levels.

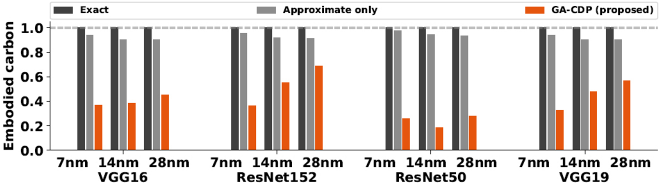


Fig. 3: Embodied carbon comparison across DNN models (Normalized to exact implementation for each technology node)

To validate our methodology, we evaluated VGG16, VGG19, ResNet50, and ResNet152, on ImageNET subset, across 7, 14, and 28 nm nodes. Figure 3 compares three designs: the exact baseline meeting a 30 FPS threshold, an approximate version using area-efficient multipliers with up to 2.0% accuracy drop, and our proposed solution (GA-CDP). While approximation alone reduces embodied carbon, our methodoogy further optimizes the design through minimal architecture and efficient multipliers. Results show significant reductions in embodied carbon across all networks and nodes, with up to 65% savings for VGG16 and 30%–70% for others, proving that our approach creates sustainable, performance-compliant designs.

IV. Conclusion

We proposed a carbon-aware DNN accelerator design using approximate computing and architecture exploration. Our approach achieves up to 70% lower carbon footprint with minimal accuracy loss, meeting performance requirements across various models and technology nodes.

Acknowledgments

This work has been supported by grant NSF CCF 2324854.

References

[1] U. Gupta *et al.*, "Chasing carbon: The elusive environmental footprint of computing," in *2021 IEEE HPCA*, 2021.

[2] A. M. Panteleaki *et al.*, "Carbon-aware design of dnn accelerators: Bridging performance and sustainability," in *2024 IEEE ISVLSI*.

[3] U. Gupta *et al.*, "Act: Designing sustainable computer systems with an architectural carbon modeling tool," in *ISCA*, 2022.

[4] C. C. Sudarshan *et al.*, "Eco-chip: Estimation of carbon footprint of chiplet-based architectures for sustainable vlsi," in *2024 IEEE HPCA*.

[5] K. Balaskas *et al.*, "Variability-aware approximate circuit synthesis via genetic optimization," *IEEE TCAS-I*, 2022.

[6] NVIDIA, "Nvdla primer," https://nvdla.org/primer.html, 2017.

[7] M. Gao *et al.*, "Tangram: Optimized coarse-grained dataflow for scalable nn accelerators," in *ASPLOS*, 2019.

[8] J. Gong *et al.*, "Approxtrain: Fast simulation of approximate multipliers for dnn training and inference," *IEEE TCAD*, 2023.