Carbon-Efficient 3D DNN Acceleration: Optimizing Performance and Sustainability

Aikaterini Maria Panteleaki*, Konstantinos Balaskas†, Georgios Zervakis†, Hussam Amrouch‡, Iraklis Anagnostopoulos*
*School of Electrical, Computer & Biomedical Engineering, Southern Illinois University Carbondale, USA
†Department of Computer Engineering & Informatics, University of Patras University, Greece
‡Chair of AI Processor Design; Munich Institute of Robotics and Machine Intelligence, Germany

***Abstract*—As Deep Neural Networks (DNNs) continue to drive advancements in artificial intelligence, the design of hardware accelerators faces growing concerns over embodied carbon footprint due to complex fabrication processes. 3D integration improves performance but introduces sustainability challenges, making carbon-aware optimization essential. In this work, we propose a carbon-efficient design methodology for 3D DNN accelerators, leveraging approximate computing and genetic algorithm-based design space exploration to optimize Carbon Delay Product (CDP). By integrating area-efficient approximate multipliers into Multiply-Accumulate (MAC) units, our approach effectively reduces silicon area and fabrication overhead while maintaining high computational accuracy. Experimental evaluations across three technology nodes (45nm, 14nm, and 7nm) show that our method reduces embodied carbon by up to 30% with negligible accuracy drop.**

***Index Terms*—DNN Accelerators, 3D Integration, Approximate Computing, Embodied Carbon Footprint, Sustainable Computing**

I. Introduction

The rapid growth of Artificial Intelligence (AI) has resulted in the wide adoption of Deep Neural Networks (DNNs) as a fundamental component of modern computing systems. To efficiently support the computational demands of DNNs, specialized hardware accelerators have been developed, offering significant improvements in throughput and energy efficiency. These accelerators have enabled AI deployment across a wide range of environments, from large-scale data centers to resource-constrained edge devices. However, as AI applications continue to scale in complexity, the environmental impact of DNN accelerators has become a growing concern [1]. While prior optimization efforts have focused on reducing operational energy consumption, recent studies indicate that embodied carbon, arising from semiconductor manufacturing, packaging, and fabrication processes, can contribute significantly to the total carbon footprint of AI hardware, particularly for edge devices [2], [3]. Addressing these sustainability challenges requires novel architectural strategies that balance performance with carbon efficiency.

To enhance performance and energy efficiency, many hardware designers have adopted three-dimensional (3D) integration as a promising architectural strategy for DNN accelerators. Unlike standard two-dimensional (2D) designs, 3D integration enables vertical stacking of multiple chiplets using advanced packaging techniques such as hybrid bonding and Through-Silicon Vias (TSVs) [4]. This vertical integration reduces interconnect lengths, increases memory bandwidth, and minimizes chip footprint, making it particularly beneficial for resource-constrained environments like edge and mobile computing systems. By leveraging 3D architectures, accelerators can integrate more computational and memory resources within the same physical footprint, achieving lower latency and reduced power consumption compared to their 2D counterparts [5]. However, despite these performance advantages, the manufacturing complexity of 3D integration introduces additional sustainability challenges. The embodied carbon footprint of 3D accelerators is significantly higher than that of equivalent 2D designs due to increased wafer processing steps, bonding requirements, and lower fabrication yields [6]. As a result, there exists a critical trade-off between adopting 3D architectures for their computational advantages and mitigating their environmental impact.

To address this sustainability challenge, recent methods have explored the trade-offs between performance and carbon footprint in DNN accelerator design, primarily focusing on optimizing 2D architectures [2], [7]. While these studies provide valuable insights, the transition to 3D architectures creates embodied carbon concerns due to additional manufacturing overhead. Specifically, processes such as wafer thinning, TSV etching, and bonding not only increase energy consumption during fabrication but also contribute to material waste and reduced yields [4]. This issue is particularly important in smaller accelerators designed for edge applications, where the benefits of reduced chip area may be outweighed by the higher carbon cost of 3D integration [6]. As AI accelerators support more and more mobile and embedded platforms, finding sustainable design methodologies for 3D accelerators becomes essential to balance computational efficiency with embodied carbon footprint.

Approximate computing has been widely explored as a technique for reducing energy consumption in DNN accelerators, leveraging the inherent error resilience of deep learning models [8]. Since minor inaccuracies in arithmetic computations typically have minimal impact on inference accuracy, approximate multipliers have been introduced in Multiply-Accumulate (MAC) units to achieve power and energy efficiency [9]. However, despite its success in lowering operational energy, approximate computing has not yet been explored as a strategy to mitigate embodied carbon. By replacing exact multipliers in MAC units with carefully selected approximate counterparts, it is possible to reduce silicon area, fabrication complexity, and material usage, all of which contribute directly to the embodied carbon footprint. Unlike conventional energy-focused optimizations, this approach explicitly targets the environmental cost of semiconductor manufacturing, making it particularly beneficial for 3D-integrated architectures, where embodied carbon is a dominant factor.

In this work, we present a carbon-aware design methodology for 3D DNN accelerators that leverages approximate computing to reduce embodied emissions while maintaining high computational efficiency. We explore a 3D memory-on-logic architecture, where the bottom die houses computational logic, including a systolic array of Processing Elements (PEs), while the top die integrates a global SRAM buffer for storing input activations, weights, and output feature maps [5], [10]. Since MAC units are among the most area-intensive components, we replace conventional multipliers with area-efficient approximate multipliers, significantly reducing silicon usage while maintaining acceptable accuracy for DNN inference. To systematically optimize this trade-off between performance, accuracy, and sustainability, we employ a genetic algorithm-based design space exploration framework, which searches for architectures that minimize Carbon Delay Product (CDP) while ensuring efficient computation. The key contributions

of this paper are as follows: ❶ We introduce a sustainability-driven approximation strategy that replaces exact multipliers in MAC units with optimized approximate counterparts, reducing silicon area and the associated embodied carbon footprint. ❷ We develop a multi-objective optimization framework that explores key architectural parameters, such as the number of Processing Elements (PEs), local buffer sizes, and memory hierarchy—while balancing performance, accuracy, and sustainability metrics. ❸ We define and optimize CDP as a metric that simultaneously considers both embodied carbon footprint and computational efficiency, ensuring that the resulting accelerator designs achieve a sustainable balance between performance and embodied carbon emissions.

II. Related Work

The design space exploration of DNN accelerators has traditionally focused on optimizing timing, power, and area. To improve efficiency, prior works have explored co-optimization of hardware architecture and DNN mapping strategies. In [11], a genetic algorithm-based framework optimizes hardware parameters and mapping for 2D DNN accelerators. Alternatively, [12] employs gradient descent to navigate the unified design space. Moving beyond monolithic 2D accelerators, [13] introduces a tiled accelerator paradigm, enabling both inter-layer and layer-by-layer optimizations. With the rise of 3D integration, design space exploration has expanded to multi-chiplet architectures. In [14], both Reinforcement Learning (RL) and non-RL algorithms optimize chiplet-based AI accelerators for performance, power, area, and cost, targeting Large Language Models (LLMs). The evaluation framework in [15] analyzes 2D and 3D systolic arrays, emphasizing energy consumption, while [6] extends this work by integrating sustainability analysis, assessing both operational and embodied carbon emissions.

For a holistic assessment of DNN accelerator designs, many researches have highlighted the importance of sustainable hardware approaches. The work in [3] introduces an architectural carbon footprint modeling framework that enables early-stage design space exploration. Building on this foundation, [16] investigates carbon-aware design parameters for virtual and extended reality systems, while jointly optimizing operational and embodied carbon alongside with job execution latency. However, recent studies [17] have demonstrated that embodied and operational emissions are estimated on different scales and therefore cannot be directly compared. Focusing specifically on DNN accelerators, [2] explores designs that balance the critical trade-off between sustainability and performance for 2D systolic array architectures. More advanced modeling tools presented in [18] and [19] provide detailed carbon assessments for 2.5D and 3D chips, incorporating crucial manufacturing parameters, such as technology node, yield, packaging type and most importantly the chip area. While these works provide valuable carbon modeling methodologies, our approach proposes approximate computing as a strategy to mitigate embodied carbon footprint of 3D DNN accelerators, offering a novel approach to design sustainability-aware devices with high performance.

Approximate computing leverages the inherent error tolerance of deep neural networks, as computational imprecisions in operations can have minimal impact on the overall accuracy. In the context of DNN accelerators, a wide variety of works demonstrate the benefits of approximate computing, leading to significant reductions in energy consumption, job delay and area footprint, while maintaining acceptable accuracy levels. The works in [9], [20] explore the trade-offs between energy efficiency and accuracy loss when incorporating reconfigurable approximate multipliers in the accelerator logic blocks. The authors in [21] investigate the accuracy impact of approximate

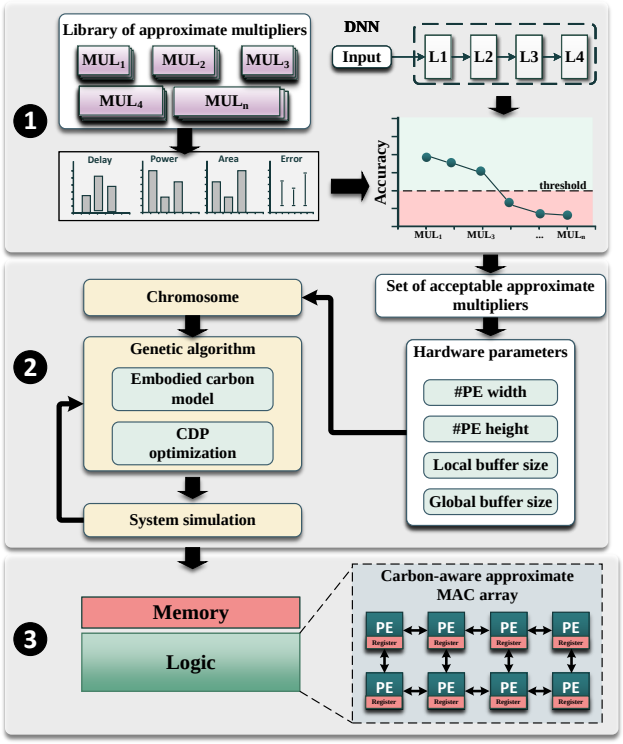


Fig. 1: Overview of the proposed methodology

circuits for various types of Convolutional Neural Networks, while the tools presented in [22], [23] offer complete frameworks that evaluate the training or inference accuracy of a DNN model when incorporating approximate multipliers. While these works highlight the gains from using approximate circuits in DNN applications, they examine only 2D accelerator architectures. Our work uniquely applies approximation in 3D DNN accelerators, as a strategy to reduce the embodied carbon footprint and address the sustainability challenges that are very common in 3D integrated systems.

III. Methodology

Figure 1 presents an overview of our proposed methodology for designing carbon-aware 3D DNN accelerators using approximate computing and a genetic algorithm-based optimization strategy. The first stage involves a library of approximate multipliers, where various multiplier designs are evaluated based on delay, power, area, and error, ensuring that only configurations meeting predefined accuracy thresholds are selected. The second stage integrates these approximate multipliers into the hardware design space exploration, where a genetic algorithm optimizes the Carbon Delay Product (CDP) by selecting the best combination of Processing Element (PE) configurations, buffer sizes, and approximate multipliers. Finally, the optimized carbon-aware 3D accelerator is evaluated, to verify the gains in embodied carbon without performance compromises.

A. 3D memory-on-logic accelerator

For our framework, we adopt a DNN accelerator model based on the Eyeriss architecture [24], which features a mesh-based Processing Element (PE) array and a hierarchical memory system optimized for deep neural network (DNN) workloads. Each PE includes a Multiply-Accumulate (MAC) unit responsible for performing core computations, as well as a local buffer in the form of a register file that stores DNN weights and intermediate results. To facilitate efficient data movement, the accelerator exchanges data with off-chip DRAM through a global

SRAM memory , which is accessible by all PEs and is used to store input activations, weights, and output feature maps.

In conventional 2D accelerator designs, data transfer between the global SRAM and the PEs relies on a Network-on-Chip (NoC), introducing communication overhead and energy inefficiencies. However, our design extends this architecture by leveraging memory-on-logic 3D integration paradigm, where the logic layer (bottom die) hosts the PE array and compute elements, while the memory layer (top die) integrates the global SRAM. This 3D stacking allows for tighter coupling between processing and memory resources, reducing data transfer latency and energy consumption.

For vertical connectivity between the two dies, we utilize hybrid bonding technology, a state-of-the-art 3D integration technique that provides higher bandwidth and lower interconnect resistance compared to traditional methods like microbumps [10]. This approach enhances performance by enabling denser and more energy-efficient inter-die communication. Additionally, the final 3D packaging employs Through-Silicon-Vias (TSVs), which establish direct electrical connections between the stacked dies and the package substrate. These TSVs further optimize data movement efficiency, which is particularly beneficial for memory-intensive DNN workloads . By minimizing off-chip memory accesses, our design significantly improves performance and energy efficiency, making it well-suited for sustainable DNN acceleration.

B. Embodied Carbon in 3D chips

The *embodied carbon footprint* represents the total CO_2 emissions associated with a device over its entire life cycle. However, for semiconductor-based systems, it is primarily dominated by emissions from fabrication processes in semiconductor manufacturing facilities [3]. In this work, we analyze the embodied carbon footprint of 3D DNN accelerators by considering emissions from the fabrication of the logic die, memory die, as well as the bonding and packaging processes that occur in fabrication plants. To quantify the embodied carbon emissions, we follow the analytical models proposed in [15], [19], which provide fine-grained estimations for 3D-integrated circuits. The total embodied carbon of a 3D DNN accelerator is formulated as:

$$C_{\text{embodied}} = C_{\text{die}}^{\text{logic}}\, C_{\text{die}}^{\text{memory}}\, C_{\text{bonding}}\, C_{\text{packaging}} \tag{1}$$

where: (i) $C_{\text{die}}^{\text{logic}}$ and $C_{\text{die}}^{\text{memory}}$ represent the carbon emissions from the fabrication of the logic and memory dies accordingly; (ii) C_{bonding} represents the emissions from 3D bonding techniques (e.g., hybrid bonding, microbumps); and (iii) $C_{\text{packaging}}$ represents the emissions from final chip packaging and interconnect integration. For each die, the fabrication-related carbon emissions are calculated as:

$$C_{\text{die}} = \text{CFPA} \times A_{\text{die}}\ \text{CFPA}_{\text{Si}} \times A_{\text{wasted}} \tag{2}$$

where: (i) CFPA represents the carbon footprint per unit area of the die; (ii) A_{die} is the effective area of the manufactured chip; (iii) CFPA_{Si} accounts for silicon wastage due to the wafer dicing process; and (iv) A_{wasted} is the unused silicon area on the wafer. The carbon footprint per unit area (CFPA) is defined as:

$$\text{CFPA} = \frac{\text{CI}_{\text{fab}} \times \text{EPA}\ \text{C}_{\text{gas}}\ \text{C}_{\text{material}}}{\text{Y}} \tag{3}$$

where: (i) CI_{fab} is the carbon intensity of the fabrication facility (depends on the energy sources used in the fab); (ii) EPA is the energy per unit area required for manufacturing; (iii) C_{gas} represents the greenhouse gas emissions from fabrication; (iv) $\text{C}_{\text{material}}$ is the carbon cost of raw material procurement; and (v) Y represents the fraction of successfully fabricated dies (yield percentage). Similarly, the embodied carbon associated with 3D bonding and packaging is proportional to the respective surface areas involved:

$$C_{\text{bonding}} = \text{CFPA}_{\text{bonding}} \times A_{\text{die}} \tag{4}$$

$$C_{\text{packaging}} = \text{CFPA}_{\text{packaging}} \times A_{\text{package}} \tag{5}$$

where $\text{CFPA}_{\text{bonding}}$ and $\text{CFPA}_{\text{packaging}}$ define the carbon cost per unit area for bonding and packaging processes, respectively.

C. Die Area Estimation for Embodied Carbon Modeling

From the above equations, it is clear that the chip area is the dominant factor affecting the embodied carbon emissions of a 3D accelerator. For the memory die, which consists of the global SRAM buffer, we utilize CACTI [25] to estimate its area based on memory capacity and technology node. In cases where CACTI does not support a specific target technology node, we apply area scaling trends from [19] to extrapolate the required values.

The logic die area, on the other hand, is determined primarily by the number of PEs in the design. Each PE contains a local buffer and a Multiply-Accumulate (MAC) unit, both of which contribute significantly to the total silicon footprint. The local buffer, implemented as a register file, follows a similar area estimation approach to that of the global SRAM, relying on existing area scaling models. However, the MAC unit is the most area-intensive component of the PE and thus has a significant impact on the overall logic die area.

To further break down the MAC unit's area composition, we analyze its arithmetic circuits under the widely used bfloat16 representation, which has been extensively adopted in modern DNN accelerators, such as Google's TPUs [26]. Each MAC unit consists of several key arithmetic blocks, including a 7-bit multiplier for mantissa multiplication, two 8-bit adders for exponent addition, and a 24-bit integer adder (23 bits mantissa plus the hidden bit) for accumulating the mantissa results. Among these components, multipliers are by far the most silicon-intensive, dominating the area requirements of the MAC unit. Given their substantial contribution to the overall die size, reducing the area of multipliers directly impacts the total embodied carbon footprint of the accelerator. This observation motivates our focus on employing approximate multipliers, as their adoption can lead to significant reductions in silicon area, thereby lowering the accelerator's carbon footprint while maintaining computational efficiency.

D. Approximate multipliers in 3D MAC units

To mitigate the embodied carbon footprint of 3D DNN accelerators, we employ *approximate computing* as a primary strategy to reduce the silicon area of the circuits that form the Multiply-Accumulate (MAC) units. Among the arithmetic components in a MAC unit, multipliers are the most resource-intensive in terms of both area and power consumption. Unlike adders, multipliers require a significantly larger number of logic gates, making them ideal candidates for approximation. Importantly, deep neural networks (DNNs) exhibit a high tolerance to approximation due to their inherent redundancy and statistical properties. The error resilience of DNN workloads arises from two key factors. First, the inference process relies more on relative weight distributions rather than precise numerical values, meaning that small deviations in computation often do not substantially alter the model's final output [8]. Second, the accumulation of multiple operations in deep architectures leads to an averaging effect, where individual approximation errors tend to cancel out rather than propagate destructively [9].

In our design, we replace conventional exact multipliers with area-efficient *approximate multipliers* from the EvoApprox library [27], which provides a wide range of arithmetic circuits with different

trade-offs between computational accuracy and area. Specifically, we target the *7-bit multipliers* responsible for *mantissa multiplication* in the bfloat16 MAC unit. By employing approximate versions of these multipliers, we significantly reduce the logic complexity, leading to lower transistor count and smaller die area. Given that the embodied carbon footprint of a die is proportional to its silicon area, this reduction translates directly into lower CO_2 emissions. To ensure that the use of approximate multipliers does not significantly lower model accuracy, we utilize *ApproxTrain* [23], a framework designed to systematically simulate the propagation of approximation errors in DNN architectures. ApproxTrain evaluates the impact of replacing exact arithmetic operations with approximate versions, allowing us to quantify the trade-offs between accuracy and area savings.

E. Genetic algorithm-based design space exploration

To systematically explore the design space of our 3D accelerator and guide the carbon optimization process, we develop a *multi-objective genetic algorithm* that efficiently navigates the vast configuration space to find optimal solutions that minimize the *Carbon Delay Product (CDP)*. This approach allows us to select accelerator configurations that effectively balance the trade-off between performance and embodied carbon footprint, a particularly challenging problem in 3D architectures due to increased fabrication complexity and energy costs. The genetic algorithm represents each accelerator configuration as a *chromosome*, encoding key hardware parameters, including the Processing Element (PE) array dimensions, local buffer size, and global SRAM capacity. Each chromosome is structured as:

$$\mathbf{C} = \{P_x, P_y, B_{\text{local}}, B_{\text{global}}\} \tag{6}$$

where: (i) P_x, P_y denote the dimensions of the PE array; (ii) B_{local} represents the local buffer size per PE; and (iii) B_{global} represents the global SRAM capacity.

Our framework integrates nn-dataflow [13] for mapping exploration, performing delay-optimized dataflow scheduling for specified hardware architectures and DNN models. We extended nn-dataflow to support memory-on-logic 3D architectures, which leverage vertical high-bandwidth interconnects between the PE array and the global SRAM. By incorporating these low-latency connections, the tool accurately models the reduced communication cost between the logic and memory layers, optimizing the dataflow strategy accordingly.

As aforementioned, our framework integrates EvoApprox approximate multipliers [27], selecting the most area-efficient multiplier that satisfies an accuracy drop constraint. For each target DNN, we impose a maximum allowable accuracy degradation of 1%, 2%, and 3%, ensuring that the accelerator maintains acceptable inference performance. Given a neural network and its associated accuracy constraint δ, the framework selects the most area-efficient approximate multiplier M_{approx} that satisfies:

$$\Delta AM_{\text{approx}} \leq \delta \tag{7}$$

where $\Delta AM_{\text{approx}}$ represents the accuracy drop introduced by using a specific approximate multiplier M_{approx} compared to an exact multiplier.

The genetic algorithm follows a structured evolutionary process to explore the accelerator design space. Before the optimization process, *ApproxTrain simulations* [23] are used to evaluate the accuracy impact of different approximate multipliers. Only multipliers that satisfy the predefined accuracy constraints (Eq. 7) are included in the design space. In *Step 1: Initialization*, a population of N candidate configurations (chromosomes) is randomly generated. In *Step 2: Fitness Evaluation*, each candidate is assessed based on multiple criteria: the *carbon model* (Eq. 1) is used to compute C_{embodied}, and the *nn-dataflow performance model* estimates D_{task}. In *Step 3: Selection*, the top-performing designs with the lowest CDP values are chosen for reproduction. In *Step 4: Crossover*, new candidate configurations are generated by recombining parameter values from selected solutions. To introduce diversity and prevent premature convergence, *Step 5: Mutation* applies random modifications to some parameters. Finally, in *Step 6: Termination*, the algorithm iterates for G generations until convergence criteria are met, ensuring that the best accelerator designs are identified based on both computational efficiency and embodied carbon reduction.

IV. Evaluation

We evaluated our framework considering inference delay and embodied carbon footprint across three technology nodes: 45nm, 14nm, and 7nm. These nodes were selected to show scalability and applicability to modern fabrication processes. The 45nm node, used in the EvoApprox library [27], provides post-synthesis area data for approximate multipliers. For 14nm and 7nm, we synthesized the approximate multipliers using the Synopsys Design Compiler to obtain accurate area estimations. The clock frequencies for each node are scaled appropriately with existing technology advancements and set to 500 MHz (45nm), 940 MHz (14nm), and 1050 MHz (7nm).

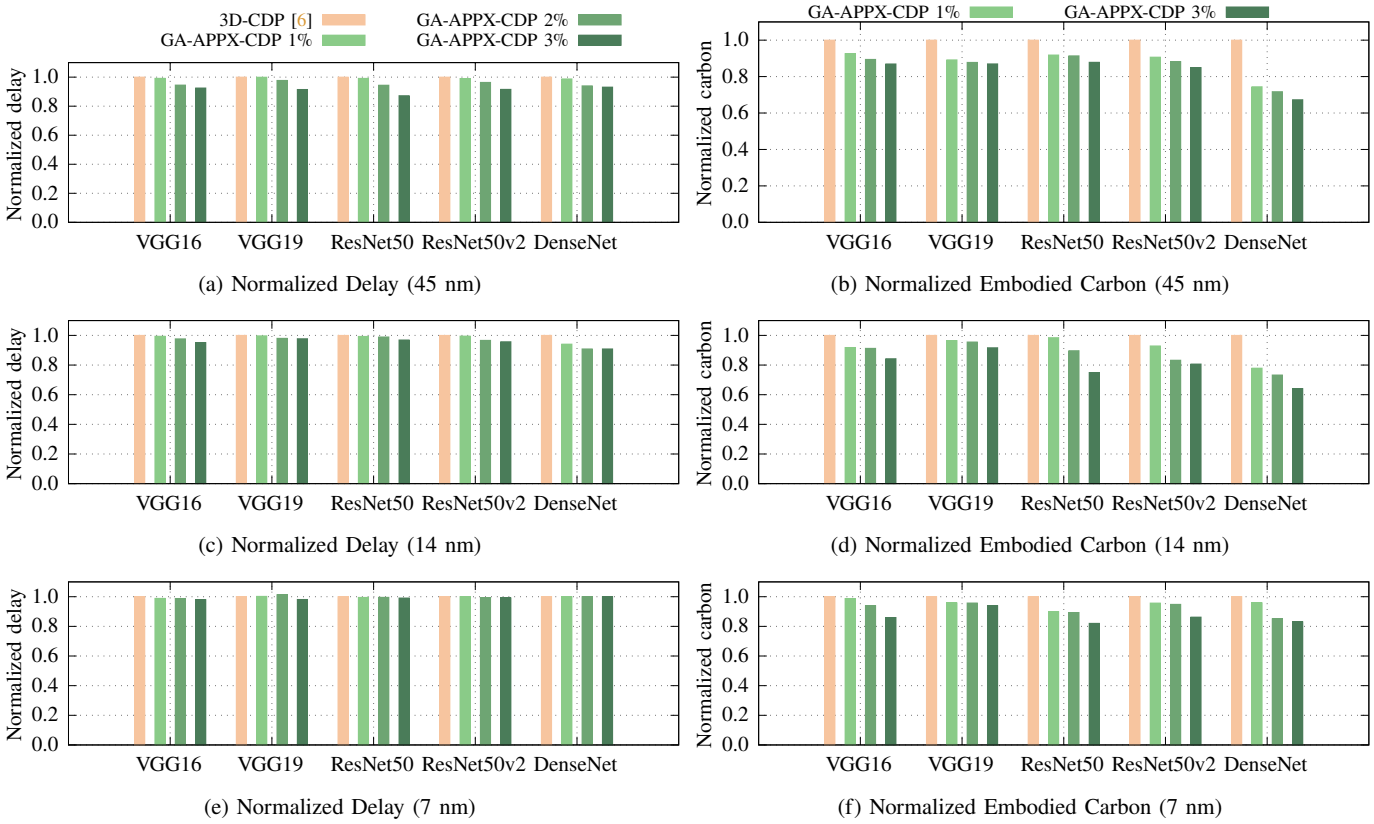
To evaluate the impact of the approximate multipliers on DNN inference accuracy, we integrated their netlists into ApproxTrain and tested them on ImageNet dataset. We selected five widely used CNNs, VGG16, VGG19, ResNet50, ResNet50V2, and DenseNet, due to their architectural diversity and utilization in edge computing applications, particularly in AR/VR workloads where resource efficiency is critical. To ensure a balance between accuracy and carbon reduction, we considered three accuracy drop thresholds: 1%, 2%, and 3%, selecting only the approximate multipliers that maintained accuracy within these limits each time.

For embodied carbon estimation, we utilized detailed manufacturing parameters from [10], which models a 3D DNN System-on-Chip incorporating hybrid bonding for inter-chiplet connections, TSV packaging, and Face-to-Face wafer stacking. The Carbon Footprint Per Area (CFPA), area scaling trends, and yield values for different technology nodes are derived from the methodologies presented in [18], [19], providing a comprehensive foundation for accurate carbon footprint assessment across the different selected fabrication processes.

A. Delay and Embodied Carbon comparison

In Figure 2, we show the results in terms of normalized delay and normalized carbon for all three selected technology nodes (45nm, 14nm, and 7nm) and three accuracy drop thresholds (1%, 2%, and 3%). As a baseline, we used the optimization approach presented in [6], which explores CDP optimization for 3D DNN accelerators without utilizing approximate computing techniques. Our results demonstrate that the integration of approximate multipliers along with the genetic algorithm-based design space exploration consistently reduces the embodied carbon footprint across all configurations while maintaining competitive performance.

The most significant reductions are observed at 45nm and 14nm, where our framework, *labeled as GA-APPX-CDP*, achieves up to 25% and 30% lower embodied carbon, respectively, compared to the baseline. At 7nm, the improvements are more limited, reaching only 15%, as the relative impact of silicon area reductions diminishes in smaller process nodes due to increased interconnect overhead and static power dissipation. The benefits of approximate computing are particularly evident in compute-bound models, such as DenseNet, where carbon reductions reach 30%, driven by a fundamental shift



(a) Normalized Delay (45 nm)

(b) Normalized Embodied Carbon (45 nm)

(c) Normalized Delay (14 nm)

(d) Normalized Embodied Carbon (14 nm)

(e) Normalized Delay (7 nm)

(f) Normalized Embodied Carbon (7 nm)

Fig. 2: Comparison of normalized inference delay and embodied carbon across different technology nodes (45nm, 14nm, and 7nm). GA-APPX-CDP is our proposed method for different accuracy drop thresholds.

in the design space exploration process. By reducing MAC unit area, the genetic algorithm favors configurations with higher PE counts, increasing parallelism and reducing inference latency while maintaining efficient workload distribution. This results in a positive feedback loop, where additional PEs not only accelerate computations but also improve sustainability by distributing processing more efficiently. In contrast, exact computing baselines require larger, more carbon-intensive units, limiting their ability to optimize both performance and carbon footprint simultaneously. The relationship between accuracy drop thresholds (1%, 2%, and 3%) and embodied carbon follows a clear trend, where higher tolerance for approximation leads to greater area savings, directly reducing the carbon footprint. These results show that approximate computing enables a co-optimization of computational efficiency and sustainability, making it a highly effective strategy for 3D-integrated DNN accelerators.

B. Performance-constrained carbon optimization

Previous studies [3] have shown that modern DNN accelerators are often overdesigned, delivering more performance than needed for edge applications while significantly increasing their embodied carbon footprint. To analyze the trade-off between carbon efficiency and performance, we modeled NVDLA-like architectures with PE arrays ranging from 64 to 2048 PEs, scaling in powers of two. The local and global buffer sizes were proportionally adjusted based on MAC array dimensions, following NVIDIA's specifications [28].

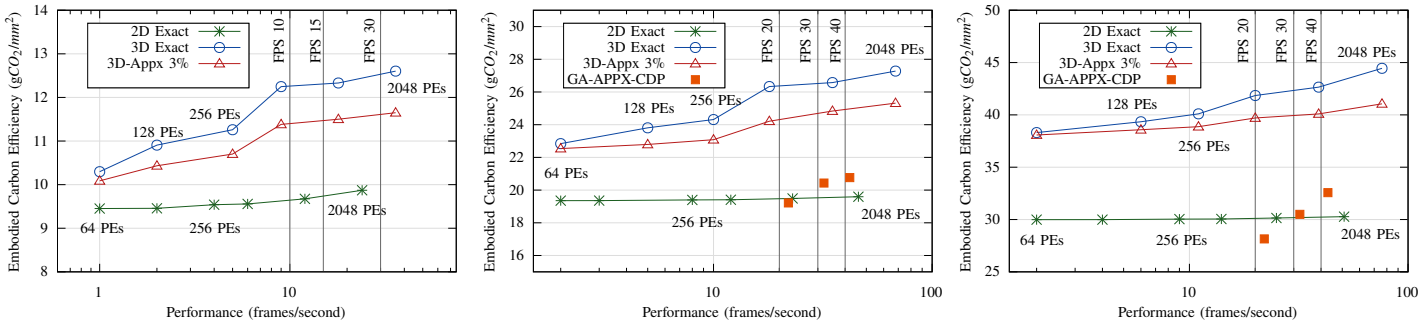
Our evaluation compares four design approaches: (1) 2D architectures with exact multipliers (2D Exact); (2) 3D architectures with exact multipliers (3D Exact); (3) 3D architectures with approximate multipliers (3D-Appx), which allow up to 3% inference accuracy degradation; and (4) our genetic algorithm-optimized solution (GA-APPX-CDP), which minimizes CDP while meeting realistic FPS constraints. The GA-APPX-CDP optimization ensures that the final design satisfies accuracy drop thresholds and practical edge system requirements. According to the technology node, we applied FPS targets of 10, 15, 20, 30, and 40 to guide our optimization process.

Figure 3 shows that conventional 3D accelerators consistently outperform 2D designs in computational performance due to vertical integration and reduced interconnect latency. However, this comes at a significant carbon cost, as 3D designs exhibit higher embodied carbon per unit area. Integrating approximate multipliers reduces this overhead, lowering the carbon footprint of 3D accelerators while maintaining competitive performance. Despite these improvements, the carbon gap between 3D and 2D architectures remains notable, reflecting the manufacturing costs of 3D stacking. Our GA-APPX-CDP solutions achieve a better balance, retaining the performance benefits of 3D integration while approaching the carbon efficiency of 2D designs. This advantage is particularly prevalent in advanced technology nodes, such as 14nm and 7nm, where our approach meets target FPS thresholds of 20, 30, and 40 with significantly better carbon efficiency than conventional 3D implementations. For instance, at 7nm with a 20 FPS target, our solution achieves 32% better carbon efficiency compared to an exact 3D accelerator with similar performance and 7% lower carbon per unit area than a 2D architecture satisfying the same FPS threshold.

Due to space limitations, we present results only for VGG16, but the observed trends are consistent across all evaluated DNN models.

V. Conclusion

This work shows that approximate computing and genetic algorithm-based design space exploration effectively reduce the embodied carbon



(a) 45nm (b) 14nm (c) 7nm

Fig. 3: Embodied carbon efficiency (gCO_2/mm^2) and performance (frames/second) for VGG16 across different technology nodes. The x-axis is in logarithmic scale. GA-APPX-CDP is our proposed method, which minimizes CDP while meeting realistic FPS constraints.

footprint of 3D DNN accelerators while preserving high computational efficiency. Our results show that carbon reductions of up to 30% at 14nm and 25% at 45nm can be achieved by integrating approximate multipliers, with accuracy degradation up to 3%. Moreover, at 7nm with a 20 FPS constraint, our optimized designs improve carbon efficiency by 32% compared to exact 3D implementations and lower carbon per unit area by 7% relative to 2D designs. These findings confirm that a co-optimized approach, combining approximate computing and CDP-driven optimization, enables sustainable and efficient AI hardware.

Acknowledgments

This work is supported by grant NSF 2324854. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] U. Gupta *et al.*, "Chasing carbon: The elusive environmental footprint of computing," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 854–867.

[2] A. M. Panteleaki and I. Anagnostopoulos, "Carbon-aware design of dnn accelerators: Bridging performance and sustainability," in *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 2024.

[3] U. Gupta *et al.*, "Act: Designing sustainable computer systems with an architectural carbon modeling tool," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, 2022, pp. 784–799.

[4] S. Jeloka *et al.*, "System technology co-optimization and design challenges for 3d ic," in *2022 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2022, pp. 1–6.

[5] L. Yang *et al.*, "Three-dimensional stacked neural network accelerator architectures for ar/vr applications," *IEEE Micro*, 2022.

[6] H. J. Byun *et al.*, "Energy-/carbon-aware evaluation and optimization of 3d ic architecture with digital compute-in-memory designs," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, 2024.

[7] A. M. Panteleaki *et al.*, "Late breaking results: Leveraging approximate computing for carbon-aware dnn accelerators," *arXiv preprint arXiv:2502.00286*, 2025.

[8] Z.-G. Tasoulas *et al.*, "Weight-oriented approximation for energy-efficient neural network inference accelerators," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4670–4683, 2020.

[9] O. Spantidi *et al.*, "Positive/negative approximate multipliers for dnn accelerators," in *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 2021, pp. 1–9.

[10] T. F. Wu *et al.*, "11.2 a 3d integrated prototype system-on-chip for augmented reality applications using face-to-face wafer bonded 7nm logic at $< 2\mu$m pitch with up to 40% energy reduction at iso-area footprint," in *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 67. IEEE, 2024, pp. 210–212.

[11] S.-C. Kao *et al.*, "Digamma: Domain-aware genetic algorithm for hw-mapping co-optimization for dnn accelerators," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2022.

[12] C. Hong *et al.*, "Dosa: Differentiable model-based one-loop search for dnn accelerators," in *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, 2023, pp. 209–224.

[13] M. Gao *et al.*, "Tangram: Optimized coarse-grained dataflow for scalable nn accelerators," in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 807–820.

[14] K. Mishty and M. Sadi, "Chiplet-gym: Optimizing chiplet-based ai accelerator design with reinforcement learning," *IEEE Transactions on Computers*, 2024.

[15] H. J. Byun *et al.*, "3d ic architecture evaluation and optimization with digital compute-in-memory designs," in *Proceedings of the 29th ACM/IEEE International Symposium on Low Power Electronics and Design*, 2024.

[16] M. Elgamal *et al.*, "Design space exploration and optimization for carbon-efficient extended reality systems," *arXiv preprint arXiv:2305.01831*, 2023.

[17] N. Bashir *et al.*, "On the promise and pitfalls of optimizing embodied carbon," *ACM SIGENERGY Energy Informatics Review*, vol. 4, no. 3, pp. 94–99, 2024.

[18] Y. Zhao *et al.*, "3d-carbon: An analytical carbon modeling tool for 3d and 2.5 d integrated circuits," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024, pp. 1–6.

[19] C. C. Sudarshan *et al.*, "Eco-chip: Estimation of carbon footprint of chiplet-based architectures for sustainable vlsi," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2024, pp. 671–685.

[20] O. Spantidi and I. Anagnostopoulos, "The perfect match: Selecting approximate multipliers for energy-efficient neural network inference," in *2023 IEEE 24th International Conference on High Performance Switching and Routing (HPSR)*. IEEE, 2023, pp. 27–32.

[21] G. Zervakis *et al.*, "Leveraging highly approximated multipliers in dnn inference," *IEEE Access*, 2025.

[22] F. Vaverka *et al.*, "Tfapprox: Towards a fast emulation of dnn approximate hardware accelerators on gpu," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2020, pp. 294–297.

[23] J. Gong *et al.*, "Approxtrain: Fast simulation of approximate multipliers for dnn training and inference," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[24] Y.-H. Chen *et al.*, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE journal of solid-state circuits*, vol. 52, no. 1, pp. 127–138, 2016.

[25] H. Packard, "Cacti: An integrated cache and memory access time, cycle time, area, leakage, and dynamic power model," GitHub repository, available: https://github.com/HewlettPackard/cacti.

[26] S. Wang and P. Kanwar, "Bfloat16: The secret to high performance on cloud tpus," Google Cloud Blog, 2019. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus

[27] V. Mrazek *et al.*, "Evoapprox8b: Library of approximate adders and multipliers for circuit design and benchmarking of approximation methods," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. IEEE, 2017.

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