SGen: Sustainable Generative AI Inference Using Chiplet-based In-Memory Computing Architectures

Tarun Sharma, Deepank Grover, Naorem Akshaykumar, Sujay Deb
IIIT Delhi

Abstract

Generative AI (GenAI) models deployed on monolithic chips face scalability, power, and environmental challenges. As the usage of GenAI grows, its high carbon footprint poses environmental concerns. In this work, we propose SGen, a framework for sustainable inference using chiplet-based in-memory computing (IMC) architectures. SGen selects mapping strategies dynamically and schedules the inference based on the availability of renewable energy resources and latency requirements. By utilizing chiplet modularity, 3D integration, and carbon-aware scheduling, SGen significantly lowers environmental impact, making GenAI more sustainable at scale.

Keywords

Generative AI, Chiplets, Sustainability, In-memory compute (IMC)

1 Introduction

Generative AI (GenAI) is rapidly expanding across various computational domains and is expected to be a center for a wide range of emerging applications. However, as sequence length grows to meet ever-escalating user demands, the corresponding GenAI models scale in complexity and size, increasing the resource requirements for deployment. Realizing such large GenAI models on conventional monolithic silicon incurs prohibitive area overheads, reduced fabrication yield, and limits scalability. To overcome these architectural bottlenecks, chiplet-based design paradigms are being pursued for GenAI workloads [8]. The use of chiplets introduces modularity and scalability, enabling hardware platforms to adapt more flexibly to increasing model dimensions. Moreover, the reduced die size of chiplets enhances manufacturing yield and significantly shortens time-to-market. Chiplet-based systems also demonstrate a lower carbon footprint (CFP) compared to their monolithic counterparts, making them indispensable for sustainable GenAI [11].

An emerging and critical dimension of GenAI is its impact on the environment. The conventional reliance on power-hungry GPUs during GenAI inference incurs substantial energy consumption and cooling demands [13]. As GenAI systems become deeply embedded in daily life, the carbon emissions associated with it assume higher importance. The environmental footprint of GenAI has two categories: embodied CFP, which arises from the materials, chemicals, and energy expended during integrated circuit fabrication; and operational CFP, which accounts for the emissions generated during runtime execution. While chiplet-based architectures offer a substantial reduction in embodied CFP due to their smaller, modular design, addressing the environmental cost of GenAI at scale demands a concerted effort to minimize both embodied and operational CFP.

A promising strategy for sustainable GenAI is the deployment of IMC within chiplet-based architectures, commonly referred to as chiplet-based IMC systems [5]. By reducing data movement between memory and compute units, IMC offers an energy-efficient and high-throughput alternative to traditional von Neumann architectures. Coupled with reduced silicon area requirements, IMC is more sustainable than conventional hardware paradigms [2]. The adoption of 3D integration in chiplet-based IMC architectures further improves performance by minimizing interconnect latency compared to 2.5D implementations [12]. However, the increased spatial density of chiplets in 3D stacks imposes thermal management challenges, as the limited vertical heat dissipation introduces localized hotspots [9]. When mapping GenAI workloads onto such thermally constrained architectures, rising temperatures increase cooling demands. Therefore, there is a need to co-optimize latency and operational CFP in the architectural design of chiplet-based IMC systems for GenAI deployment.

With the above factors in mind, we propose *SGen* for sustainable GenAI inference for chiplet-based IMC architectures. *SGen* takes the type of inference and type of power source (renewable/non-renewable) into account and schedules the GenAI model to reduce operational CFP incurred during inference. It also dynamically decides the mapping strategy. Here, the mapping strategy refers to the mapping of the layers of the GenAI model with the IMC tiles. We quantify the effectiveness of SGen using a Vision Transformer [3] with the Imagenet dataset, containing over 86M parameters.

2 Background

2.1 Online and Offline Inference

The inference for GenAI models can be classified into online and offline inference. In online inference, the model is expected to generate outputs immediately in response to user inputs. This inference is used in applications like chatbots and search suggestions. Offline inference, meanwhile, refers to the processing of inference requests that are not time-sensitive and can tolerate higher latency. Using it allows for the accumulation and batch processing of requests. Offline inference can constitute up to 55% of the total workload in some cases [6]. Overall, while the latency of the system is critical in online inference, it is not a major factor in offline inference.

2.2 Mapping Strategies

In this work, we consider two mapping strategies. The first mapping strategy (Mapping 1) uses the DNN weight and mapping method [5]. The top-to-bottom approach allocates weights onto the IMC architecture's hardware resources, structured hierarchically into tiers, tiles, processing elements, and crossbars. This approach is aimed at reducing the inter-chiplet data movement, which results

'CogArch 2025', June 22, 2025, Tokyo, Japan
2025. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

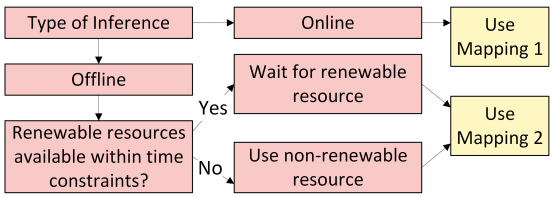

Figure 1: Block diagram of Carbon Aware Scheduling

in higher performance. A side effect of this mapping strategy is the high intra-chiplet movement. While this mapping reduces the inter-chiplet movement, the higher intra-chiplet movement increases the peak temperature. This is more pronounced at the lowest layer of the chiplet. The second mapping strategy (Mapping 2) uses a thermal-aware task mapping algorithm [10]. This strategy aims to reduce the peak temperature at the cost of mapping at lower performance. This reduces the operational carbon incurred for cooling. Therefore, while Mapping 1 focuses on performance, Mapping 2 focuses on operational carbon.

3 Sustainable Inference using Carbon Aware Scheduling

Fig. 1 shows the block diagram of the proposed carbon-aware scheduling. During the process, the scheduler first determines the kind of inference. For latency-sensitive online inference, the processor uses Mapping 1 to reduce the latency incurred during inference. The mapping is done irrespective of the kind of energy source available during the process. This ensures that the critical tasks of users are not delayed. However, this results in high power consumption due to the cooling requirements of the server.

For offline inference with relaxed latency requirements, the scheduler first checks for renewable energy resource availability within the time constraints of the inference. It aims to schedule the batches during periods of peak renewable energy availability to reduce operational CFP. For instance, wind energy generation often peaks during the late afternoon to evening hours [7]. Additionally, the scheduler uses Mapping 2, which has a lower peak temperature compared to Mapping 1. The use of both renewable energy and thermal aware mapping reduces operational CFP considerably. However, it may happen that sufficient renewable resources are not available within the required timing constraints for offline inference. In that case, the scheduler uses only Mapping 2 with the existing non-renewable resources. While the operational CFP might not reduce a lot due to the use of non-renewable resources, it is still significant. This is because the offline inference is performed multiple times over the chip's lifetime. Therefore, even a slight reduction in operational CFP leads to a noticeable cumulative impact.

4 Results

We demonstrate our results for a vision transformer using the ImageNet dataset. We first obtain the power, energy, and latency using the HISIM simulator [12]. We choose a configuration of 3D packaging with 3 tiers. Each tier contains 49 tiles, and each tile contains 36 processing elements. Each processing elements consist of an RRAM array of size 1024×1024 with IMC infrastructure. Table 1 shows the various parameters of the two mapping algorithms for

Mapping	Latency (μs)	Power (W)	Energy (mJ)	Peak Temp (K)
Mapping 1	15.8	3.21	35.73	367.77
Mapping 2	17.4	2.92	34.62	313.95

Table 1: Output parameters for mapping strategies

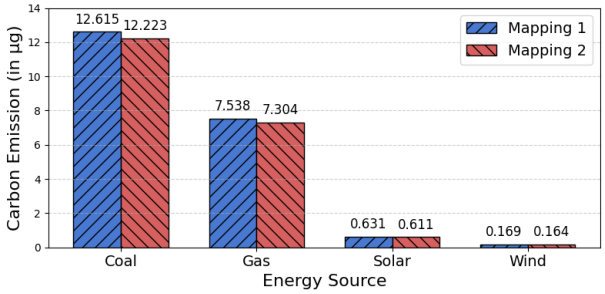

Figure 2: Operational CFP for various sources

one inference of the vision transformer. Mapping 1 exhibits lower latency, while Mapping 2 shows lower power, energy, and peak temperature.

To calculate the operational CFP of the server, both the computational power and the power required for cooling must be considered. As of 2022, the average power usage effectiveness of the IT industry stood at 1.55 [1]. This indicates that for every joule of energy consumed by the server, an additional 0.55 J of energy is consumed for cooling only. Thus, we scale the server energy values reported in Table 1 by a factor of 1.55 to account for total energy consumption. After converting the energy consumed into kilowatt hours, we use the carbon intensity of the electricity source to obtain the operational CFP. The values of carbon intensity of energy sources are obtained from the work ACT [4].

The values of operational CFP for various sources are shown in Fig. 2. The value of operational CFP is obtained for a single inference of the vision transformer. When SGen operates using renewable energy, the reduction in operational CFP ranges from 76.9× (from coal to wind) to 12.3× (from gas to solar). When SGen relies on fallback using non-renewable energy, the reduction in operational CFP for both coal-based and gas-based inference is 3.11%. This difference in operational CFP is less due to the smaller difference in energy between the two mapping strategies (in order of mJ). The higher latency of Mapping 2 counters the power savings obtained from it, reducing the difference between it and Mapping 1. More efficient mapping strategies are the need of the hour to enable sustainable inference.

5 Conclusion

Generative AI's environmental impact poses a challenge as models grow in complexity. Addressing both performance and CFP is essential to ensure the long-term viability of the models. In this work, we introduced SGen, a carbon-aware inference framework that combines chiplet-based IMC architectures with carbon-aware scheduling and mapping strategies. By considering latency sensitivity and energy source availability, SGen successfully balances computational efficiency with environmental responsibility. Future works can explore more sustainable mapping strategies for GenAI inference.

References

[1] BOYD. 2024. *Energy Consumption in Data Centers: Air versus Liquid Cooling.* Retrieved April 2, 2025 from https://www.boydcorp.com/blog/energy-consumption-in-data-centers-air-versus-liquid-cooling.html

[2] Hyung Joon Byun, Udit Gupta, and Jae-sun Seo. 2024. 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* (Newport Beach, CA, USA) *(ISLPED '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3665314.3670838

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[4] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. 2022. ACT: designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture* (New York, New York) *(ISCA '22)*. Association for Computing Machinery, New York, NY, USA, 784–799. https://doi.org/10.1145/3470496.3527408

[5] Gokul Krishnan, Sumit K. Mandal, Chaitali Chakrabarti, Jae-Sun Seo, Umit Y. Ogras, and Yu Cao. 2021. System-Level Benchmarking of Chiplet-based IMC Architectures for Deep Neural Network Acceleration. In *2021 IEEE 14th International Conference on ASIC (ASICON)*. 1–4. https://doi.org/10.1109/ASICON52560.2021.9620238

[6] Yueying Li, Zhanqiu Hu, Esha Choukse, Rodrigo Fonseca, G. Edward Suh, and Udit Gupta. 2025. EcoServe: Designing Carbon-Aware AI Inference Systems. arXiv:2502.05043 [cs.DC] https://arxiv.org/abs/2502.05043

[7] Yang Liu and Jie Bai. 2023. Daily Variation and Regional Differences in Wind Power Output during Heat and Cold Wave Days in China. *International Transactions on Electrical Energy Systems* 2023, 1 (2023), 8828093. https://doi.org/10.1155/2023/8828093 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1155/2023/8828093

[8] Mohanad Odema, Luke Chen, Hyoukjun Kwon, and Mohammad Abdullah Al Faruque. 2024. SCAR: Scheduling Multi-Model AI Workloads on Heterogeneous Multi-Chiplet Module Accelerators. In *2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 565–579. https://doi.org/10.1109/MICRO61859.2024.00049

[9] Shailja Pandey, Lokesh Siddhu, and Preeti Ranjan Panda. 2023. NeuroCool: Dynamic Thermal Management of 3D DRAM for Deep Neural Networks through Customized Prefetching. *ACM Trans. Des. Autom. Electron. Syst.* 29, 1, Article 19 (Dec. 2023), 35 pages. https://doi.org/10.1145/3630012

[10] Lili Shen, Ning Wu, Gaizhen Yan, and Fen Ge. 2017. Thermal-aware task mapping for communication energy minimization on 3D NoC. *IEICE Electronics Express* 14, 22 (2017), 20170900–20170900. https://doi.org/10.1587/elex.14.20170900

[11] Chetan Choppali Sudarshan, Nikhil Matkar, Sarma Vrudhula, Sachin S. Sapatnekar, and Vidya A. Chhabria. 2024. ECO-CHIP: Estimation of Carbon Footprint of Chiplet-based Architectures for Sustainable VLSI. In *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 671–685. https://doi.org/10.1109/HPCA57654.2024.00058

[12] Zhenyu Wang, Pragnya Sudershan Nalla, Jingbo Sun, A. Alper Goksoy, Sumit K. Mandal, Jae-sun Seo, Vidya A. Chhabria, Jeff Zhang, Chaitali Chakrabarti, Umit Y. Ogras, and Yu Cao. 2025. HISIM: Analytical Performance Modeling and Design Space Exploration of 2.5D/3D Integration for AI Computing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2025), 1–1. https://doi.org/10.1109/TCAD.2025.3531348

[13] Adam Zewe. 2025. *Explained: Generative AI's environmental impact.* Retrieved April 2, 2025 from https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117