# Amortizing Embodied Carbon Across Generations

Shixin Ji*, Jinming Zhuang*, Zhuoping Yang*, Alex K. Jones†, Peipei Zhou*
* Brown University, † Syracuse University
{shixin_ji, jinming_zhuang, zhuoping_yang, peipei_zhou}@brown.edu, akj@syr.edu

***Abstract*—Data centers have been relying on renewable energy integration coupled with energy efficient specialized processing units and accelerators to increase sustainability. Unfortunately, the carbon generated from manufacturing these systems is becoming increasingly relevant due to these energy decarbonization and efficiency improvements. Furthermore, it is less clear how to mitigate this aspect of embodied carbon. As workloads continue to evolve over each hardware generation we explore the tradeoffs of fabricating new application-tuned hardware compared with more general solutions such as Field Programmable Gate Arrays (FPGAs). We also explore how REFRESH FPGAs can amortize embodied carbon investments from previous generations to meet the requirements of future generations workloads.**

***Index Terms*—sustainable computing, embodied carbon, cloud, data center, hardware accelerators, machine learning**

## I. Introduction

The focus on operational energy efficiency for computing systems over the last decade has led to significant advancements in sustainability in data centers. Unfortunately, some of these advancements have come without attention to and sometimes through exacerbation of the carbon from manufacturing computing systems. This manufacturing carbon, the principle source of *embodied* carbon, is much more difficult to reduce as it is dominated by semiconductor integrated circuits, and to a lesser extent, other necessary electronics to build these systems. While these processes can also benefit from renewable energy integration, there are fundamental aspects of the process from raw element extraction to biproducts of the fabrication process itself that limit the mitigation potential of embodied carbon from manufacturing [1]–[3].

To address this problem there has been increasing interest in extending the lifetimes of computing hardware to amortize the embodied carbon over a longer timescale [4]. A relatively simple solution that is widely discussed is replacing chip-based accelerators with Field Programmable Gate Arrays (FPGAs) [5]. Thus, one investment in an FPGA allows the hardware to be optimized and updated to support new applications as they appear and evolve without incurring further embodied carbon.

A challenge with this approach is that FPGAs do evolve over time to better serve application needs. For instance, in the context of machine learning, recently we have seen an evolution from convolutional networks, to transformers, and, more recently, to Generative Pre-Trained Transformers (GPT). However, the trend does not pressure increases in computational power, data size, and access bandwidth equally. For instance, convolutional networks depend heavily on computation speed, while transformers moderately increase this need for computation they much more heavily depend on larger working sets for their computation. GPT's and large language models further increase the need for data while the computational requirement only increases moderately.

Previous work noted that FPGA architecture has remained relatively stagnant for many generations, with an emphasis on increasing the effective hardware real-estate [6]. While many of the fundamental claims that the conceptual architectures do track, the impact of larger and faster multiply-accumulate units, larger and more efficient on-chip data memory, improved access to off-chip memory, and the biggest architectural change of devices with dedicated tensor cores can have a big impact on how FPGAs perform with emerging workloads [7]–[10].

In this paper we propose *gradual embodied carbon investing* using chiplets through 2.5D integration. We call these embodied carbon optimized (ECO) systems. Presuming a capability to open and modify a package [6], this technique builds a system around FPGA hardware tuned to solve emerging algorithms with increasingly large workloads. Consider two approaches, the traditional approach one is to provision a large FPGA device when creating a new system. This device is considerably larger than is required to solve the typical workload at the time of provisioning. This allows the system to grow and expand over time to adapt to and execute new workloads with larger problem sizes and computational requirements. This is the premise of many hyperscalar vendors and is characterized to show it is more sustainable than directly fabricating accelerator chips [5].

In contrast, an ECO-FPGA would contain a device sized to accelerate common workloads at that time without considerable additional size. Immediately this is more efficient as the FPGA is now sized to the problem size and will be more performance and energy-efficient. However, this FPGA die would be connected through an interposer which had open spots to add additional devices over time. As workloads adapt over time it is expected that additional FPGA resources would be required additional dies be added in the package to supplement the existing hardware. The value proposition is that increases in embodied carbon from introducing an interposer and from the process of adding additional dies to the system are far outweighed by the benefit of introducing new hardware tuned to newly developed workloads. Moreover, many systems already leverage chiplet techniques to improve die yields normalizing the cost of adding in an interposer.

Thus, an ECO-FPGA grows over time with newly released FPGA hardware while still having the ability to leverage the older hardware through reconfigurability. The proposition is that an ECO-FPGA can replicate the capability of a single release FPGA with a largely reduced embodied carbon cost.

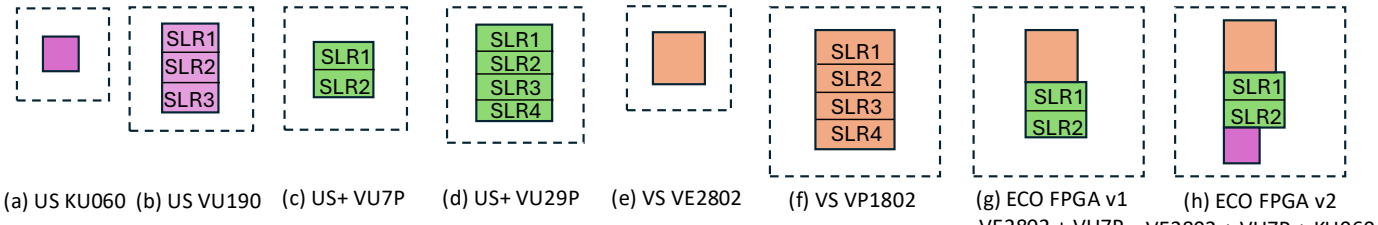


Fig. 1: FPGA layouts.

TABLE I: FPGA family, part, year released, and workloads.

| Line | FPGA | Part | Year | Workload |
|---|---|---|---|---|
| 1 | US Kintex | XCKU060 | 2015 | AlexNet |
| 2 | US Virtex | XCVU190 | 2015 | |
| 3 | US+ Virtex | XCVU7P | 2018 | Resnet-152 (16-bit) |
| 4 | US+ Virtex | XCVU29P | 2018 | |
| 5 | VS VEK280 | VE2802 | 2021 | ViT-Base (INT8) |
| 6 | VS VPK180 | VP1802 | 2021 | |

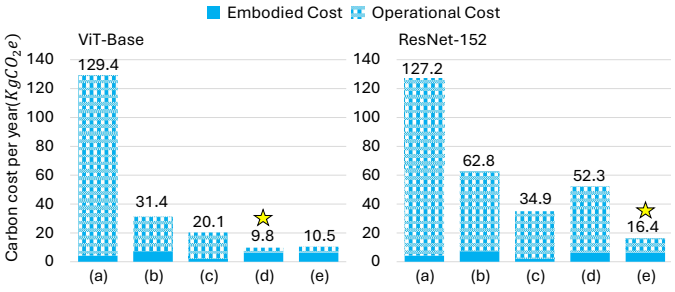


Fig. 2: Yearly carbon cost consumptions of (a) XCVU29P (2018), (b) VP1802 (2021), (c) VE2802 (2021), (d) ECO-FPGA v1, (e) ECO-FPGA v2, for ViT-Base in INT8 and ResNet-152 in INT16. The number of each device is set to 1 and the throughput is set to the same as one XCVU29P. The carbon intensity value is set to $0.188\,\mathrm{kg\,CO_2e}$, which is the intensity in NY state, US [11].

## II. Case Studies

To demonstrate the concept of ECO-FPGAs, we consider two case studies, one built over the course of three generations Xilinx/AMD Ultrascale (US), Ultrascale+ (US+), and Versal (VS), released in 2015, 2018, and 2021, respectively. The other case study considers just US+ and VS from 2018, and 2021, respectively. Each of these generations was released to support critical workloads of that generation such as convolutional neural networks (CNNs) the size of AlexNet in 2015; continued use of AlexNet and introducing more complex CNNs such as ResNet-152 in 2018; and continued use of complex CNNs like ResNet-152 with high-precision transformers like ViT-Base (INT8) in 2021. We summarize the menu of chiplets and the emerging workloads in Table I.

To illustrate the design of the various FPGA concepts we studied we illustrate them in Fig. 1. We start with the two generation case study, denoted ECO-FPGA v1. To implement the ResNet-152 or INT-8 ViTBase would be possible with a US+ XCVU7P (Fig. 1-c). Following the GreenFPGA [5] model would allocate an XCVU29P for headroom in the next generations (Fig. 1-d). In 2021 we would add a VS VE2802 chiplet (Fig. 1-e) to form the ECO-FPGA v1 (Fig. 1-g).

We show the carbon cost of running the ViT-Base workload that was emerging in 2021 as well as continued use of ResNet-152 in Fig. 2. We use the ACT [2] tool to quantify the embodied costs and compute operational carbon using performance and energy statistics of the FPGA coupled with a carbon intensity factor of $0.188\,\mathrm{kg\,CO_2e}$. The XCVU29P is not well suited to the transformer because of its relatively low computational capability, resulting in considerable energy-inefficiency substantial operational carbon. The VEK280 alone is much better than the XCVU29P due to the AI Engine but it still has a substantial carbon cost from both operational overhead and embodied carbon. The operational overhead is due to insufficient on-chip storage 51MB to store the weights which imposes substantial off-chip memory access.

ECO-FPGA v1 has only a minor embodied carbon increase over a new VE2802 part alone, but reduces operational carbon substantially by combining the better compute engine with sufficient on-chip storage making it the overall better choice. Interestingly, a larger versal part VP1802 (Fig. 1-f) is even less suitable because it has higher embodied carbon from additional chip area but still lower energy efficiency due to using FPGA fabric instead of using AI Engine for the tensor computation.

The other case study is to consider an ECO-FPGA started in 2015. In this case, the target workload would have been AlexNet, which could have been satisfied using a Kintex part, but using the GreenFPGA model might have been replaced by a larger Virtex part for headroom (lines 1–2 of Table I). The trend for adding in a KU060 part in 2018 for AlexNet and comparable CNNs is similar to the trend in ECO-FPGA v1 for ResNet-152 in 2021, with the core data not shown due to space limitations. However, we do show the ECO FPGA v2 that includes chiplets from line 1,3,5 of Table I in Fig. 2. The US part coupled with the others remains useful and improves the performance of ResNet-152 while maintaining a similar performance for the transformer.

## III. Conclusions

In this paper we propose ECO-FPGA, and embodied carbon optimized approach to add-in smaller FPGA chiplets from multiple generations, while amortizing the embodied carbon compared to provisioning larger FPGAs. Our results show that by leveraging resources targeted towards the evolution of workloads provides better results compared to replacement with similar size next generation devices or deploying large older generation devices with extra headroom to grow.

## REFERENCES


[1] D. Kline, N. Parshook, X. Ge, E. Brunvand, R. Melhem, P. K. Chrysanthis, and A. K. Jones, "Holistically evaluating the environmental impacts in modern computing systems," in *2016 Seventh International Green and Sustainable Computing Conference (IGSC)*. IEEE, 2016, pp. 1–8.

[2] U. Gupta, M. Elgamal, G. Hills, G.-Y. Wei, H.-H. S. Lee, D. Brooks, and C.-J. Wu, "ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool," in *Proc. of ISCA*, 2022, p. 784–799. [Online]. Available: https://doi.org/10.1145/3470496.3527408

[3] S. Ji, Z. Yang, X. Chen, S. Cahoon, J. Hu, Y. Shi, A. K. Jones, and P. Zhou, "SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators," in *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2024, pp. 496–501.

[4] I. Samaye, P. Leloup, G. Sassatelli, and A. Gamatié, "Towards Sustainable Low Carbon Emission Mini Data Centres," 2024. [Online]. Available: https://arxiv.org/abs/2405.01909

[5] C. C. Sudarshan, A. Arora, and V. A. Chhabria, "GreenFPGA: Evaluating FPGAs as environmentally sustainable computing solutions," 2023.

[6] P. Zhou, J. Zhuang, S. Cahoon, Y. Tang, Z. Yang, X. Chen, Y. Shi, J. Hu, and A. K. Jones, "REFRESH FPGAs: Sustainable FPGA chiplet architectures," in *Proc. of IGSC*, 2023, p. 1–3.

[7] J. Zhuang, J. Lau, H. Ye, Z. Yang, Y. Du, J. Lo, K. Denolf, S. Neuendorffer, A. Jones, J. Hu, D. Chen, J. Cong, and P. Zhou, "CHARM: Composing Heterogeneous AcceleRators for Matrix Multiply on Versal ACAP Architecture," in *The 2023 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3543622.3573210

[8] Z. Yang, J. Zhuang, J. Yin, C. Yu, A. K. Jones, and P. Zhou, "AIM: Accelerating Arbitrary-precision Integer Multiplication on Heterogeneous Reconfigurable Computing Platform Versal ACAP," in *ICCAD*, 2023.

[9] P. Dong, J. Zhuang, Z. Yang, S. Ji, Y. Li, D. Xu, H. Huang, J. Hu, A. Jones, Y. Shi, Y. Wang, and P. Zhou, "EQ-ViT: Algorithm-Hardware Co-Design for End-to-End Acceleration of Real-Time Vision Transformer Inference on Versal ACAP Architecture," *IEEE TCAD*, 2024.

[10] J. Zhuang, Z. Yang, S. Ji, H. Huang, A. K. Jones, J. Hu, Y. Shi, and P. Zhou, "SSR: Spatial Sequential Hybrid Architecture for Latency Throughput Tradeoff in Transformer Acceleration," in *Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, ser. FPGA '24, 2024, p. 55–66.

[11] S. Ollivier, S. Li, Y. Tang, S. Cahoon, R. Caginalp, C. Chaudhuri, P. Zhou, X. Tang, J. Hu, and A. K. Jones, "Sustainable AI Processing at the Edge," *IEEE Micro*, vol. 43, no. 1, pp. 19–28, 2022.