EPiCarbon: A Carbon Modeling Tool for Electro-Photonic Accelerators

Farbin Fayza[1*], Cansu Demirkiran[1], Satyavolu Papa Rao[2], Darius Bunandar[3], Udit Gupta[4], Ajay Joshi[1,3]

[1]Boston University, Boston, USA [2]NY CREATES, New York, USA
[3]Lightmatter, Boston, USA [4]Cornell Tech, New York, USA

***Abstract*—The escalating carbon emissions driven by the growing computational demands of Artificial Intelligence (AI) have made energy-efficient and sustainable hardware design a high priority. Photonic computing has emerged as a promising solution, delivering orders of magnitude higher throughput and energy efficiency than CMOS for deep neural network inferences, thereby lowering operational carbon. However, studies have shown that the carbon emission from manufacturing, i.e., embodied carbon, constitutes a substantial and often dominant portion of the total carbon footprint of a computing system. Hence, it is crucial to consider both operational and embodied carbon to determine the true benefits of photonic computing. While the embodied carbon of CMOS chips and CMOS-based systems has been studied extensively, there is currently no model available for estimating the embodied carbon of photonic chips.**

In this work, we develop the first-ever model to estimate the embodied carbon of photonic chips. Our findings show that photonic chips can reduce the embodied carbon of computing systems with at least 4.1× less fabrication energy and significantly higher yield than CMOS. Building on our model, we introduce EPiCarbon, an open-source tool to evaluate the carbon footprint of Electro-Photonic (EPiC) accelerators, incorporating both operational and embodied carbon. Using EPiCarbon, we analyze the carbon footprint of state-of-the-art EPiC accelerators, demonstrating their potential as carbon-sustainable solutions for computationally demanding AI applications. Finally, through a case study on a comprehensive EPiC accelerator, ADEPT, we demonstrate key strategies to further reduce the carbon footprint of EPiC accelerators, guiding future sustainable hardware design.

***Index Terms*—photonics, machine learning, sustainability.**

I. Introduction

With the ongoing technological advancements in Artificial Intelligence (AI) and the resulting escalation in computational demands, carbon emissions from the Information and Communications Technology (ICT) sector have been rapidly increasing. Over the past decade, we have observed tremendous growth in the size and usage of AI models. The size of the GPT-based Large Language Models (LLMs) has grown by 1000× from 2018 to 2022 [1]. The rapid growth in AI usage has led to a 2.5× increase in the AI inference infrastructure at Meta within just 1.5 years [1]. Traditional CMOS-based hardware has been struggling to efficiently meet these immense demands, as it no longer scales in area or energy efficiency according to Moore's Law. Consequently, CMOS-based data centers performing trillions of daily inferences collectively emit more carbon than countries like Ireland and Denmark, with emissions projected to rise even more in the years ahead [2]–[4]. Therefore, there is an urgent need for carbon-sustainable hardware solutions to support AI's continued growth.

Photonic computing has emerged as an energy-efficient alternative to CMOS-based computing [5]–[12]. Photonics offers orders of magnitude higher throughput and energy efficiency over CMOS due to its ability to compute in the optical domain with lower losses than electrical devices [5]–[7]. Photonic computing enables efficient General Matrix-Matrix Multiplication (GEMM) operation, which constitutes over 90% of the operations in Deep Neural Networks (DNNs) [13]. As a result, researchers have designed several Electro-Photonic (EPiC) accelerators utilizing photonics for GEMM computations and electronics for the remaining computations and memory [6]–[8]. This approach shows significant advantages in terms of throughput and energy efficiency over purely electronic systems and holds promise as a sustainable computing platform for the future.

However, energy efficiency alone does not guarantee the sustainability of a technology. Recent studies have shown that the advances in energy efficiency and the increasing complexity of manufacturing have led to manufacturing (embodied) carbon becoming the dominant contributor to total carbon emissions [4], [14], [15]. As an example, embodied carbon now contributes 50%-82% of carbon emissions in cloud servers due to the improvements in energy efficiency, which reduced the operational carbon but increased the embodied carbon [16]. In the photonic computing domain, numerous studies on EPiC accelerators report orders of magnitude higher energy efficiency over CMOS-based computing [5]–[9], but none have addressed the carbon emissions associated with fabricating these accelerators. Therefore, a comprehensive study of the carbon footprint of EPiC accelerators, encompassing both operational and embodied carbon, is necessary to determine whether photonics is a viable computing solution for AI.

Several studies have analyzed the embodied carbon of CMOS chips and CMOS-based systems [14], [16]–[20], but no model currently exists for calculating the embodied carbon of photonic chips. To quantify the embodied carbon of photonic chips, we need a model that incorporates the energy consumption of each manufacturing step of photonic chips, GreenHouse Gas (GHG) emissions during fabrication, and other carbon costs. **In this work, we address this gap by developing, for the first time ever, a model to estimate the embodied**

This research was partially funded by the Institute for Global Sustainability (IGS) Graduate Student Summer Fellowship program at Boston University.

*Corresponding author: ffayza@bu.edu

carbon of photonic chips, and then use this model to analyze the carbon footprint of state-of-the-art (SOTA) EPiC accelerators, providing a comprehensive assessment of whether photonics can truly offer a carbon-sustainable hardware solution for AI. The key contributions of our work are as follows:

- We, for the first time, develop an embodied carbon model of photonic chips. We utilize the carbon cost of various processing steps [21], coupled with our estimate of a process flow to fabricate a reasonably complex photonic chip adapted from AIM-Photonics multi-project wafer technology [22], that is suitable for EPiC accelerators.
- We develop an open-source tool, EPiCarbon[1], that calculates the carbon footprint of EPiC accelerators, encompassing both operational and embodied carbon. EPiCarbon integrates our embodied carbon model for the photonic chiplets, ACT [14] for the electronic chiplets, and ECO-CHIP [17] for the heterogeneous integration of an EPiC accelerator. Using EPiCarbon, we perform a comprehensive carbon footprint evaluation of the SOTA EPiC accelerators to assess their viability as carbon-sustainable hardware for AI.
- Based on the carbon footprint evaluation of SOTA EPiC accelerators, we propose design strategies to reduce the carbon footprint of future EPiC accelerators through a case study on one of the evaluated EPIC accelerators, ADEPT [6].

Our embodied carbon model demonstrates that photonic chips require at least 4.1× less manufacturing energy than 28 nm CMOS chips and achieve significantly higher yields, resulting in lower embodied carbon than CMOS. Through the carbon footprint evaluation with EPiCarbon, we find that the SOTA EPiC accelerators outperform electronic accelerators in carbon sustainability for computationally demanding workloads due to their efficiency in both operational and embodied carbon, making them a promising solution to meet AI's growing computational needs.

II. Background and Related Work

The Carbon Footprint (CF) of a chip can be calculated with the well-established ACT model [14]. The CF of a chip has two parts: Operational Carbon Footprint (OCF) and Embodied Carbon Footprint (ECF). The total CF is the summation of OCF over the lifetime of the chip and ECF. We can also amortize ECF based on an application runtime and the lifetime of the chip to estimate the carbon footprint of an application for a specific runtime (CF_{app}) (Equation 1).

$$CF_{app} = OCF + \frac{runtime}{lifetime} \times ECF \quad (1)$$

$$OCF = Energy_{use} \times CI_{use} \quad (2)$$

$$ECF = \frac{Area}{Yield} \times (EPA \times CI_{fab} + GPA + MPA) + C_{package} \quad (3)$$

OCF is the product of the energy consumption and the carbon intensity (g/kWh) of the power source during the operation (CI_{use}) (Equation 2). ECF encompasses several components: (1) EPA (energy consumption of manufacturing per unit area of the chip) multiplied by the carbon intensity of the power source during manufacturing (CI_{fab}), (2) GPA (GHG emissions from manufacturing tools per unit area of the chip), and (3) MPA (material procurement costs per unit area of the chip) (Equation 3). Lastly, the packaging carbon cost, $C_{package}$, comes from the integration process of chiplet-based systems. ACT provides the EPA, GPA, MPA, and packaging costs for CMOS chips estimated from publicly reported data [21]. While ACT uses a fixed packaging cost, studies like ECO-CHIP [17] and 3D-Carbon [23] provide detailed estimates for different packaging types, such as 3D and 2.5D.

[1]We open-source EPiCarbon at https://github.com/bu-icsg/EPiCarbon

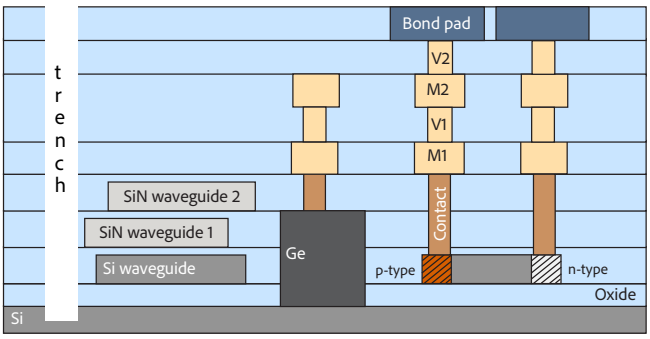

Fig. 1: Cross-section of a photonic chip adapted from AIM Photonics multi-project wafer (300 mm) technology [22].

There are also other works, such as FOCAL [19], which provides a first-order CF estimation model for processors, and GreenFPGA [18], which examines the CF of FPGAs. However, these models are not applicable to EPiC accelerators, because the ECF of photonic chips is not the same as CMOS chips due to the differences in the manufacturing process layers and yields. Our work introduces, for the first time, an ECF model for photonic chips that can be used to estimate the CF of any EPiC accelerator chip.

III. Embodied Carbon Model for Photonic Chips

We perform a detailed layer-by-layer analysis of a photonic chip's fabrication process to formulate the embodied carbon model. We use the cross-section of a photonic chip that is shown in Figure 1 with off-chip laser sources. Here, we postulate a fabrication scheme that would result in a schematic cross-section that roughly matches that shown by Fahrenkopf *et al.* [22]. The schematic cross-section is representative of a type of 'active' photonic circuit, which incorporates modulators that are necessary to manipulate the optical signals to perform computation and signal routing.

A. EPA

A significant feature of photonic chips is that they are manufactured using the same tools and, in many cases, the same processes as CMOS technology. Therefore, to estimate the EPA, we identify the necessary manufacturing steps for each layer in the chip's cross-section and use the corresponding carbon cost data from Bardon et al. [21], the same source used for CMOS EPA calculations. By adding up the carbon cost of all the steps, we get the EPA.

1) Process Layers: For the photonic chip shown in Figure 1, starting from a Silicon-on-Insulator (SOI) wafer, there are silicon (Si) and silicon-nitride (SiN) layers for fabricating the waveguides that facilitate optical signals. A germanium (Ge) photodiode layer is formed, typically by epitaxy, for sensing the processed light. The lower band-gap of germanium (compared to silicon) allows for the detection of light at a wavelength where silicon is transparent. To control the optical signals inside the Si waveguide utilizing optical computing components such as Mach-Zehnder Interferometers (MZIs) or Micro-Ring Resonators (MRRs) electro-optically, p-type and n-type regions are formed in the silicon. The Ge photodiode similarly has doped regions (forming a p-i-n diode). These are connected to heterogeneously integrated electric chips through metal layers (M1 and M2) and via layers (V1 and V2). Bond pads are placed on top to allow the chip to be connected to other chips through different heterogeneous integration methods, such as 3D or 2.5D. Finally, a deep trench is formed to expose the waveguide edge, enabling the connection of optical fibers to the photonic waveguides along the chip's edge.

2) Manufacturing Steps: In general, for all layers, the components are fabricated with the necessary lithography, etching, and cleaning steps, followed by polishing and oxide deposition by Chemical Vapor Deposition (CVD) process. The p-type and n-type regions are generated by a combination of implantation of dopant atoms into the silicon matrix (through openings in lithographically patterned photoresist) and thermal activation of the dopants. Openings (formed by reactive ion etch) on the wafer surface for metal and vias are filled with ultra-thin copper diffusion barrier and copper using a combination of deposition and metal Chemical-Mechanical Planarization (CMP) process. Additionally, the oxide that covers the waveguides is typically planarized by dielectric CMP. Lastly, the trench is created by lithography, etching, and cleaning. The full list of manufacturing steps that we have utilized in our model will be available in the EPiCarbon repository.

3) Assembling the Data and Estimating EPA: We first collect the approximate median power and throughput per wafer for each manufacturing step from the work by Bardon et al., [21]. From this data, we calculate the energy consumption for all the manufacturing steps and add them to calculate the manufacturing energy per wafer. It should be noted that different ranges in oxide thickness are utilized in photonic versus electronic chips. Thicknesses used in photonic chips are similar to the highest metal levels in CMOS chips (10th to 16th metal layers in advanced nodes) due to the need to avoid optical coupling between adjacent levels. As we use the median energy per wafer, we exclude the very low-throughput atomic layer deposition steps and very high-throughput CVD steps used for lower metal levels [21]. We also assume the facility energy to be 40% of the total energy according to [21]. In Figure 2(a), we show the EPA of a photonic chip and CMOS chips for several technology nodes. We estimate the EPA of a photonic chip to be 0.22 kWh/cm^2, which is 4.1× lower than a 28 nm CMOS chip. This difference grows with newer technology nodes because the EPA increases as the technology

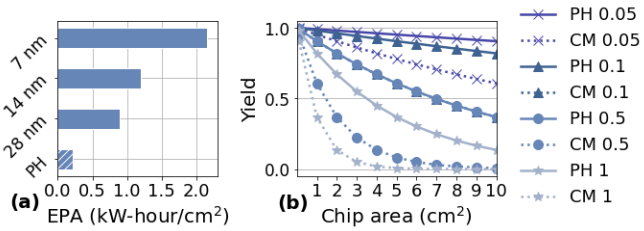

Fig. 2: Comparison of (a) EPA and (b) yield for different defect densities (cm^{-2}) of photonic (PH) and CMOS (CM) chips.

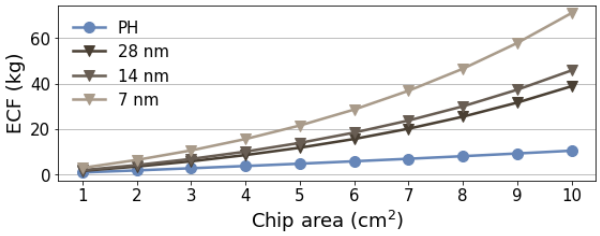

Fig. 3: Comparison of ECF growth with chip area for photonic (PH) and CMOS chips across different technology nodes (assuming coal power source and 0.1 cm^{-2} defect density). Higher yield and lower EPA than CMOS chips (Figure 2) lead to lower ECF in photonic chips.

node advances [14], [21]. As a result, the EPA of a photonic chip is 9.8× lower than a 7 nm CMOS chip.

B. GPA and MPA

We consider the GPA of a 28 nm technology node [14] as the GPA of a photonic chip. This is so that we can have a conservative estimate from publicly available data for a technology node with higher fabrication steps than photonics. In reality, it is likely to be far less because a 90 nm or 65 nm technology node can suffice for photonics (but one where 193 nm optical lithography is harnessed for resolving small spaces) [24]. We set the MPA of photonic chips as 500 g/cm^2, same as the CMOS chips [14], which is another conservative estimation due to the lack of public data for photonic chips.

C. Yield

We use the widely applied Poisson model [25] for calculating the yield of a photonic chip (as well as a CMOS chip that we use for comparison later in the paper). Photonic circuits typically use wide spaces between waveguides (to avoid optical coupling), except in portions of the circuit area associated with splitters, or points where coupling between elements is required, like a micro-ring resonator next to a waveguide. Hence, the nanometer-size defects that are typical of advanced fabs do not have the same deleterious impact on photonic chips as on densely packed CMOS chips. Taking the yield model proposed by Zhang et al., [26], the critical area can be estimated to be 20% of the total chip area. This critical area corresponds to the area occupied by the waveguides and the area around the splitters, where Zhang *et al.* [26] point out that the defect would have a significant impact. Hence, for the same defect density, photonic chips have a higher yield than CMOS by a factor of $e^{0.2}$ (Figure 2 (b)). The higher yield of photonic chips lowers the ECF by reducing the number

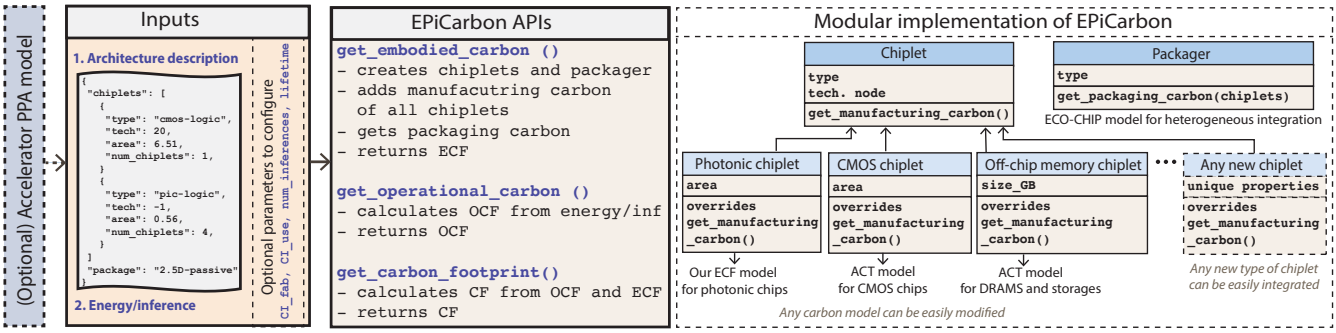

Fig. 4: Overview of EPiCarbon. The modular structure of the tool allows easy integration of any new chiplets or carbon model.

of chips wasted due to fabrication defects. As a result, with higher yield and lower EPA, photonic chips burn significantly lower embodied carbon than CMOS chips (Figure 3). In other words, we can design a large photonic chip with the same ECF as a small CMOS chip. For example, the ECF of a 6 cm^2 photonic chip is similar to a 3 cm^2 28 nm CMOS chip. Due to the impact of yield, the ECF benefits of photonic chips increase as we increase the chip area.

D. Packaging (Heterogeneous Integration)

In the SOTA EPiC accelerators, each photonic chiplet communicates with the other chiplets in the system via electrical signals. Consequently, the packaging carbon models used for fully-electric heterogeneous systems [17], [23] can be applied to heterogeneously integrated photonic and electric chips in SOTA EPiC accelerators. We use the ECO-CHIP model [17] to estimate the packaging cost in our work.

E. Correctness of the Model

To justify the correctness of our model, we make the following arguments.

- **Inclusion of all possible components:** Firstly, we consider a reasonably complex photonic chip having the necessary components to perform GEMM operations. While some components (such as SiN waveguides) of the considered photonic chip may not be required for all EPiC accelerators, we avoid underestimating the embodied carbon by considering all components.
- **Similar design across various fabrication technology:** Secondly, fabrication technologies from different sources, such as imec, amf (Singapore), or GlobalFoundries, may utilize different designs and materials or differently optimized process choices [24], [27]–[29]. We acknowledge that variations across different technologies may result in slight differences in EPA estimates. However, as supported by the references [24], [27]–[29], the overall layer structure and components across these technologies and AIMPhotonics [22] are similar, leading to EPA values that are close to ours, with variations small enough not to affect the overall architectural conclusions.
- **Validated by industry experts:** Lastly, we validate the process layers, their corresponding manufacturing steps, and the yield model by discussing them with photonics manufacturing experts from the industry. For the carbon cost of the manufacturing steps, our model uses conservative estimates of the data from imec [21], which has been widely used for embodied carbon evaluation of CMOS-based systems [4], [14], [17], [18], [23].

Therefore, our model provides a reliable and conservative estimate of the embodied carbon for photonic chips based on publicly available data.

IV. The EPiCarbon Tool

We develop the EPiCarbon tool to evaluate the CF of EPiC accelerators. Figure 4 shows an overview of the tool. The tool includes three separate APIs (Application Programming Interface) to compute the ECF, OCF, and CF individually.

Inputs: To estimate ECF, the user inputs an architecture description file in JSON format that lists the specifics for all chiplets within the accelerator. Each chiplet entry includes the chiplet type (CMOS or photonic), area, technology node (for CMOS), and quantity. Additionally, the user specifies a packaging type (3D, 2.5D-active/passive, etc.). To estimate OCF, the user needs to input the energy per inference of the accelerator. For CF, the user needs to provide both the architecture description and energy per inference. Optionally, a Power-Performance-Area (PPA) model of the accelerator (generic such as [30], [31] or accelerator-specific) can be plugged in to generate these inputs and call the desired API from EPiCarbon. We use this approach for our case study on ADEPT in Section VI.

CF Estimation: For ECF, EPiCarbon first parses the architecture description file and instantiates the corresponding chiplets and a packager. Each type of chiplet implements its own ECF model. Photonic chiplets implement our ECF model (described in Section III), whereas CMOS chiplets implement the ACT ECF model. The packager uses the ECO-CHIP model to calculate the packaging carbon based on packaging type and chiplet information. The ECF is calculated by summing the embodied carbon of all chiplets and packaging carbon. EPiCarbon also estimates the embodied carbon of off-chip memory devices using ACT if the user specifies any in the architecture description.

For OCF, along with providing the accelerator's energy per inference, the user can also configure the number of inferences per day and the lifetime of the accelerator. EPiCarbon calculates OCF for running the specified number of inferences per day over the entire lifetime of the accelerator. The tool outputs

CF by summing ECF and OCF. Alternatively, the CF for a specific runtime can be calculated using Equation 1.

Modularity of the tool: We design EPiCarbon in a modular manner by creating separate classes for each chiplet type. A user can easily include new types of chiplets by adding a new chiplet class and implementing its embodied carbon model. This makes EPiCarbon readily extendable to support other emerging technologies like phase change memory or resistive RAMs in the future.

V. CF Analysis of EPiC Accelerators

In this section, we revisit the literature on EPiC accelerators with a focus on carbon sustainability. We deploy EPiCarbon to evaluate the CF of four widely recognized EPiC accelerators along with the electronic accelerators that were used as comparison points in the corresponding EPiC accelerator papers (listed in Table I). We use the area and energy per inference numbers reported in the respective papers to compute the ECF and OCF, respectively. The reported energy numbers of the EPiC accelerators include laser, optical component tuning, data modulation, electrical-to-optical and optical-to-electrical conversions, digital-to-analog and analog-to-digital conversions, and memory accesses. The details of the energy breakdowns can be found in the respective papers. ADEPT [6], Albireo [7], and DEAP-CNN [32] run Convolutional Neural Networks (CNNs): Alexnet and VGG-16; Lightening Transformer (LT) [9] runs transformer models: BERT-large and DeiT-b.

It is worth noting that Nvidia-A100—an industry-scale accelerator with multiple functionalities, including training and sparsity optimizations—has a significantly larger area than the academic accelerators. Directly comparing the embodied carbon of A100 with other academic accelerators is not entirely fair. Therefore, we only consider the portion of the area responsible for mixed precision inference in A100 from the chip snapshot [33]. In addition to the accelerators that were compared in the EPiC accelerator papers, we include TPUv4i, an industry-scale inference accelerator. We use the reported 400 mm^2 area, 138 TFLOP/s (bf16/8b int) peak throughput, and 75 W mean power from [34], [35] in our analysis. Table II lists the EPiCarbon parameter setup that we use for our analysis, with the selected values representing typical configurations in real-world scenarios.

TABLE I: Accelerators evaluated in our work. Configuration details can be found in the corresponding accelerator papers. For EPiC accelerators, the technology node refers to the technology used for CMOS components within the accelerator.

Type	Accelerator	Configuration	Tech. node of CMOS (nm)
EPiC (Academic)	ADEPT [6]	128×128 single photocore	22
	ALBIREO [7]	ALBIREO-C: 9 PLCGs	22
	DEAP-CNN [8]	Two convolutional units	22
	Lightening-Transformer (LT) [9]	LT-L-8	14
Electronic (Academic)	Envision [36]	Fixed, as per reference	28
	UNPU [37]	Fixed, as per reference	65
Electronic (Industry)	Nvidia-A100 GPU [38]	Fixed, as per reference	7
	TPUv4i [34]	Fixed, as per reference	7

TABLE II: EPiCarbon configuration used in our work.

CI_{fab} source	Non-renewable energy (coal)
CI_{use} source	Renewable energy (wind)
Yield model	Poisson model
Defect density	0.1 cm^{-2}
Packaging type	3D (EPiC), Monolithic (electronic)
Lifetime	5 years

A. Analyzing ECF

We begin by analyzing the ECF of the SOTA EPiC accelerators compared to the electronic accelerators. Figures 5(a) and 5(b) present the reported chip area and peak throughput (TFLOP/s) of the accelerators, respectively. As these accelerators have different computing capabilities and we cannot make a head-to-head ECF comparison, in Figures 5(c) and 5(d) we show the area/TFLOP/s and ECF/TFLOP/s, respectively.

In Figure 5(c), we observe that the EPiC accelerators DEAP-CNN and LT have similar or lower area/TFLOP/s than all electronic accelerators, whereas ADEPT and Albireo have higher area/TFLOP/s than TPUv4i and A100. However, in Figure 5(d), we see that all EPiC accelerators have similar or lower ECF/TFLOP/s than the electronic ones. This is due to the lower fabrication cost and higher yield of the photonic chiplet in the EPiC accelerators, which helps to reduce the overall ECF of the chip, despite having large areas. To better understand the contribution of photonic components of the SOTA accelerators in ECF reduction, we provide a detailed breakdown of both area and ECF across the various components of the EPiC accelerators in Figure 6.

Firstly, in both ADEPT and LT, the primary contributors to their ECFs are the CMOS components, as shown in Figure 6(a). This is due to a large fraction of their area being

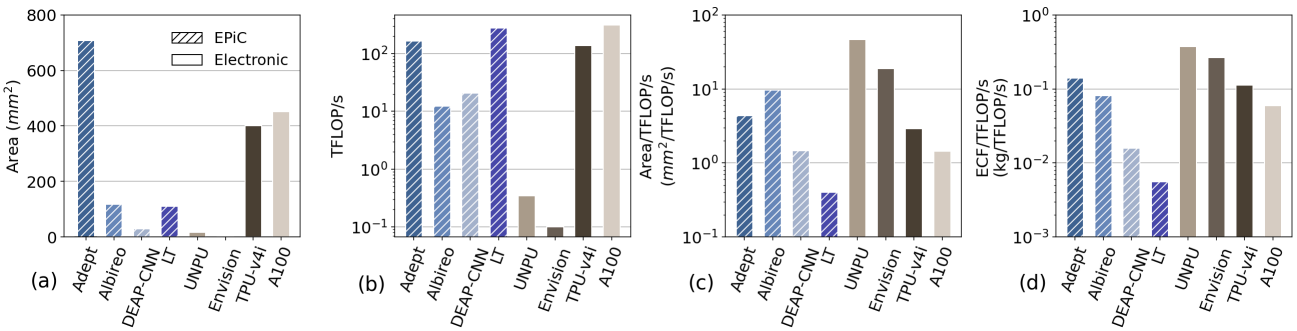

Fig. 5: (a) Area (b) peak TFLOP/s, (c) area/TFLOP/s, and (d) ECF/TFLOP/s of the accelerators. EPiC accelerators achieve similar or lower ECF/TFLOP/s despite some of them having larger area/TFLOP/s than electronic accelerators.

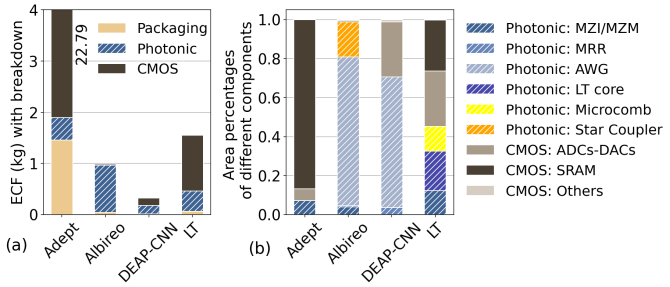

Fig. 6: (a) ECF and (b) area breakdown of the accelerators. For some EPiC accelerators, the area is largely occupied by CMOS components, leading to high ECF. Increasing the percentage of the photonic area in an EPiC accelerator helps to reduce ECF.

dedicated to the supporting electrical components. Especially in ADEPT, over 85% of the total area is occupied by on-chip SRAM, significantly increasing its ECF and resulting in the highest ECF/FLOP/s among the EPiC accelerators. As explained in Section III, CMOS components contribute more to the ECF than photonic components of the same area, as photonics provides a higher yield and lower EPA. Consequently, although the area of LT is similar to Albireo (Figure 5(a)), LT consumes ≈50% more ECF due to its higher proportion of CMOS area than Albireo. Increasing the proportion of photonic area can significantly reduce the ECF of an EPiC accelerator. However, we cannot perform all operations in the photonic domain. Thus, choosing the right ratio of photonic to electrical area in EPiC accelerators is critical to minimizing the ECF while achieving target performance.

For the cases where the photonic chiplet is the dominant source of ECF (Albireo and DEAP-CNN), it is primarily due to the large optical components like AWGs, star couplers, and microcombs [7], [9], which occupy significantly larger areas (100s-1000s μm length) compared to CMOS components. For example, in Albireo, 72% of the area is occupied by only 9 AWGs, which are necessary for efficient data distribution in its photonic compute cores [7]. Notably, DEAP-CNN is the most compact among the EPiC accelerators, with an area of ≈30 mm^2, due to its area-efficient MRR-bank-based design, resulting in the lowest ECF among the EPiC accelerators.

B. Analyzing OCF

Figure 7 shows the throughput, energy/inference/s, and the corresponding OCF/inference/s of the accelerators for the reported DNN models. As mentioned earlier in this section, the energy values are directly taken from the corresponding papers, and for EPiC accelerators, the energy includes all key components such as laser, optical component tuning, data modulation, electrical-to-optical and optical-to-electrical conversions, digital-to-analog and analog-to-digital conversions, and memory accesses.

In general, the figure demonstrates the reported advantages of EPiC accelerators, achieving orders of magnitude higher throughput and better energy efficiency than the electronic accelerators. This operational efficiency comes from fast and energy-efficient compute capabilities of photonics with lower losses than CMOS. In Figure 7(b), ADEPT consumes the

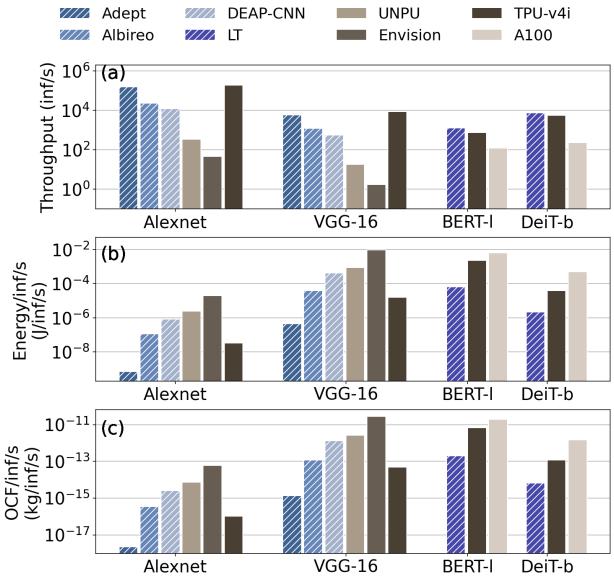

Fig. 7: (a) Throughput, (b) energy/inference/s, and (c) OCF/inference/s of the accelerators for the reported DNN models. As reported in the literature, EPiC accelerators generally provide higher throughput with better energy efficiency than electronic accelerators, leading them to consume lower OCF.

lowest energy/inference/s for AlexNet and VGG-16, while LT consumes the lowest for BERT-l and DeiT-b. For some EPiC accelerators, such as DEAP-CNN and Albireo, high data conversion costs between analog and digital domains and tuning costs of the optical devices lead to higher energy/inference/s than TPUv4i (also to keep in mind that the throughput for TPUv4i is calculated assuming maximum utilization, which is unlikely to be achieved in practice). In contrast, accelerators like ADEPT and LT have optimized tuning strategies that lead to low energy consumption. In Figure 7(c), we show OCF/inference/s with ≈3μg carbon emission per Joule.

C. Analyzing CF

Figure 8 shows the total CF (ECF + OCF) for executing one trillion inferences under varying throughput targets. One can also experiment with different total number of inferences and observe similar results. To meet these throughput goals, we scale the number of accelerator chips accordingly. As the total work (running 1T inferences) is fixed, increasing the throughput target along the X-axis shifts the CF toward being more ECF-dominant due to the increasing number of chips.

We observe that EPiC accelerators achieve lower CF than the electronic accelerators for all throughput goals, with an exception for Alexnet. This CF advantage of the EPiC accelerators comes from both OCF and ECF efficiency of photonic chips (lower carbon per throughput than CMOS). As an example, for BERT-l and DeiT-b, LT achieves the lowest CF across all throughput goals. For low throughputs achievable with a single chip, LT wins with its lowest OCF per throughput and ECF per throughput (see sections V-A and V-B for reference). At higher throughputs, where ECF dominates, its ECF efficiency becomes the key driver of its carbon advantage over the electronic accelerators.

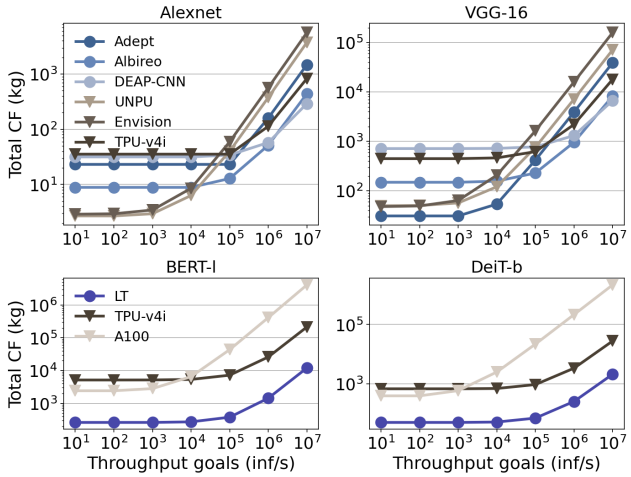

Fig. 8: Total CF of accelerators running 1 trillion inferences across different throughput goals. The SOTA EPiC accelerators are more carbon efficient than the electronic accelerators for computationally demanding workloads.

For VGG-16, ADEPT, with its highest OCF efficiency, achieves the lowest CF in the OCF dominant scenarios (up to 10^4 inference/s). However, as the throughput goal increases and the scenario becomes ECF-dominant, DEAP-CNN and Albireo become more carbon sustainable options with their higher ECF efficiency than ADEPT.

Alexnet requires significantly less computation compared to the other models—for instance, it involves only 0.7 billion FLOPs, while VGG-16 requires 15 billion FLOPs [39]. As a result, even at low throughput targets ($\leq 10^4$ inferences/s), the scenario remains ECF-dominant, and a single tiny chip like UNPU or Envision is sufficient to meet the throughput demand. This represents an edge-device-like workload, for which the current EPiC accelerators (designed for heavy computation) are over-provisioned. In such scenarios, their low utilization leads to inefficiency, making the compact, low-power electronic accelerators like UNPU or Envision a more carbon-efficient choice.

Takeaway: EPiC accelerators are more carbon-sustainable than electronic accelerators for computationally heavy workloads with efficiency in both embodied and operational carbon, making them a strong candidate for supporting the escalating compute demands of AI.

VI. Strategies to Improve CF of EPiC Accelerators: A Case Study using ADEPT

In this section, we explore several strategies to improve the CF of EPiC accelerators through an ADEPT design case study. To estimate ADEPT's CF under varying parameters, we use ADEPT's PPA estimator model developed by Demirkiran et al. [6]. and plug it in with EPiCarbon. Although ADEPT has numerous design parameters, we focus on weight SRAM size, activation SRAM size, photonic core size, and technology node of the CMOS chiplet. This is because (1) ADEPT has a large area that leads to the highest ECF among the EPiC accelerators, and the stated parameters have the highest impact on the area in ADEPT. (2) These parameters are common in all

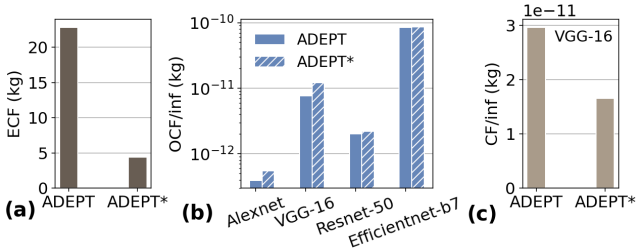

Fig. 9: (a) ECF, (b) OCF/inference, (c) CF/inference (amortized ECF) for ADEPT and ADEPT* (ADEPT with reduced weight SRAM size).

EPiC accelerators, and insights from their CF impact analysis can be broadly applied. In the following experiments, we change one parameter at a time, keeping all other parameters fixed as in the original paper (300 MB weight SRAM, 100 MB activation SRAM, 128×128 single photonic core, 22 nm CMOS chiplet, and 10 GHz frequency) [6].

A. Weight SRAM size

ADEPT has a huge 300 MB weight SRAM to minimize the cost of loading weights from the DRAM. In Figure 9, we explore the carbon impact of reducing the weight SRAM size to 32 KB (ADEPT* design), which can hold twice the number of weights that the ADEPT compute core can process at a time. We use a double buffering approach, which is a common practice in many architectures. We assume the DRAM has sufficient bandwidth (e.g., HBM with >1 TBPS bandwidth [40]) to sustain peak throughput of the accelerator. **Therefore, the throughput is not impacted by this change in the weight SRAM size.**

Reducing the size of the weight SRAM to 32KB decreases the CMOS area by 2.9× in ADEPT, resulting in a 5.2× reduction in ECF (see Figure 9(a)). However, reducing the weight SRAM size forces the model weights to be fetched from the DRAM for each batch, as the entire model can no longer fit in the SRAM. This increases the total energy consumption and, in turn, increases OCF by 1.27× on average (over four DNN workloads) (see Figure 9(b)). This OCF increase is minimal for models with higher arithmetic intensity, such as Efficientnet-b7, with only a 1.01× increase, whereas it is the highest in VGG-16 (1.6×), as VGG-16 has the lowest arithmetic intensity. Despite the increase in OCF, reducing weight SRAM size can significantly reduce CF. In Figure 9(c)), we show the CF per inference (amortized ECF for single inference time following Equation 1) for VGG-16, the model with the highest OCF increase. By reducing the weight SRAM to our proposed size, ADEPT's CF/inference is reduced by ≈44% for VGG-16, and ≈49% on average for all models.

B. Activation SRAM size

Figure 10 shows the impact of activation SRAM size on throughput, ECF, OCF/inference, and CF/inference (amortized ECF for single inference time). Increasing activation SRAM size allows for larger batch sizes, which increases throughput until the compute core reaches maximum utilization (Figure 10(a)). Larger batch sizes also reduce OCF by amortizing the weight loading cost from SRAM to photocore (Figure

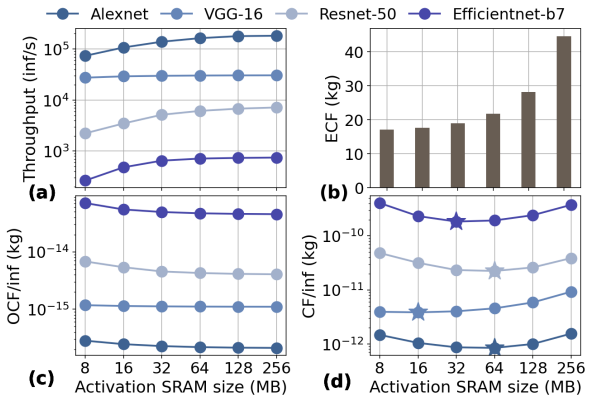

Fig. 10: Impact of the activation SRAM size on (a) throughput, (b) ECF, (c) OCF/inference, and (d) CF/inference (amortized ECF). CF-optimal activation SRAM sizes are marked with * in (d).

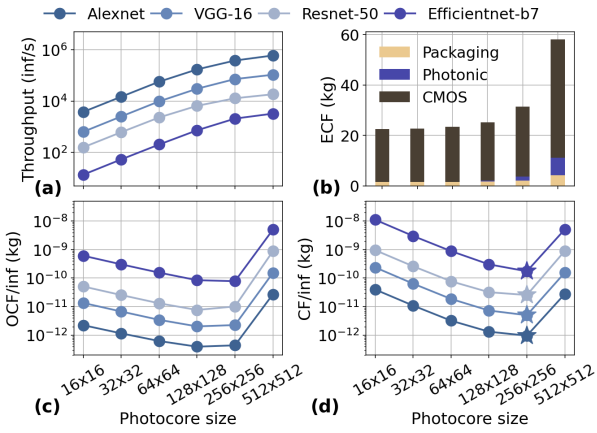

Fig. 11: Impact of the photocore size on (a) throughput, (b) ECF, (c) OCF/inference, and (d) CF/inference (amortized ECF).

10(c)). At the same time, increasing the activation SRAM size increases the chip area, raising ECF (Figure 10(b)). Thus, for a given DNN workload, there exists an optimal activation SRAM size that balances the trade-off between ECF and OCF. We have marked this SRAM size for the different networks in Figure 10(d). Here, by reducing the activation SRAM from 100 MB in the original design to 64 MB, CF/inference can be reduced by ≈18% in ADEPT on average for the evaluated workloads, without any loss in throughput.

C. Photonic core size

In Figure 11, we show the impact of different photonic core (photocore) sizes on throughput, ECF, OCF/inference, and CF/inference (amortized ECF for single inference time). Firstly, as we increase the photocore size, ECF increases (Figure 11(b)). From the ECF breakdown, we observe that although the larger photocore area contributes to the ECF increase, the CMOS ECF rises even more, dominating the total ECF. This is because to support the increased photocore size, additional CMOS components (e.g. data converters) are needed. The additional CMOS components, combined with the SRAM, significantly degrade yield, leading to a high ECF.

In Figure 11(c), we observe that OCF/inference decreases up to a certain photocore size, and then starts to increase. This

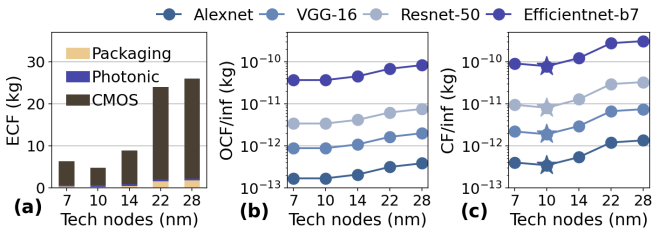

Fig. 12: Impact of technology node of CMOS on (a) ECF, (b) OCF/inference, and (c) CF/inference (amortized ECF). Going from 28 nm to up to 10 nm can significantly reduce CF/inference.

is because of the saturating throughput (Figure 11(a)), which fails to compensate for the exponentially rising laser power with the increasing photocore size [6]. As a result, in Figure 11(d), CF/inference also decreases and then starts increasing back up. For the DNN workloads in our experiment, we identify 256×256 to be the carbon-optimal photocore size, as against the 128×128 proposed in the paper, reducing ADEPT's CF/inference by ≈30% on average.

D. Technology node of CMOS chiplet

Typically, as the technology node of CMOS shrinks, it reduces energy consumption and area, which offsets the increased manufacturing carbon associated with the smaller nodes [19]. Therefore, switching to smaller nodes can reduce CF without any loss in throughput. However, advanced CMOS nodes (<10 nm) no longer shrink as per Moore's law and cannot offset the increased manufacturing carbon, which leads to diminishing CF improvements or even higher CF [19]. Hence, careful selection of the technology node for the CMOS chiplet is crucial to achieving the optimal CF in EPiC accelerators.

To evaluate the carbon impact of using different technology nodes in the CMOS chiplet of ADEPT, we use the area and energy scaling model from ECO-CHIP [17]. Figure 12 shows the change in ECF, OCF/inference, and CF/inference (amortized ECF for single inference time) across various technology nodes from 28 nm to 7 nm. Figure 12(a) shows that as the technology node scales down from 28 nm to up to 10 nm, ECF decreases significantly; however, scaling beyond 10 nm leads to an increase in ECF. The OCF/inference also reduces with smaller nodes, but the rate of reduction saturates at the advanced nodes (Figure 12(b)). As a result, in Figure 12(c), CF/inference reduces for up to 10 nm and then starts to increase. For the original ADEPT design with a 22 nm CMOS chiplet, switching to 10 nm can reduce the CF/inference by 71% on average across the DNN workloads.

VII. Conclusion

In this paper, we demonstrate the potential of photonics to offer a sustainable computing solution for DNNs, even under a conservative assessment. We present our open-source EPiCarbon tool, to estimate the CF of EPiC accelerators and perform a comprehensive CF analysis of the SOTA EPiC accelerators. We hope that this tool will benefit the computer architecture community towards developing a carbon-sustainable computing environment for AI with photonics.

REFERENCES

[1] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai *et al.*, "Sustainable ai: Environmental implications, challenges and opportunities," *Proceedings of Machine Learning and Systems*, vol. 4, pp. 795–813, 2022.

[2] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020.

[3] "Our world in data," https://ourworldindata.org/energy, 2020.

[4] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu, "Chasing carbon: The elusive environmental footprint of computing," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 854–867.

[5] S. Ning, H. Zhu, C. Feng, J. Gu, Z. Jiang, Z. Ying, J. Midkiff, S. Jain, M. H. Hlaing, D. Z. Pan *et al.*, "Photonic-electronic integrated circuits for high-performance computing and ai accelerators," *Journal of Lightwave Technology*, 2024.

[6] C. Demirkiran *et al.*, "An electro-photonic system for accelerating deep neural networks," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 19, no. 4, pp. 1–31, 2023.

[7] K. Shiflett *et al.*, "Albireo: Energy-efficient acceleration of convolutional neural networks via silicon photonics," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 860–873.

[8] J. Peng *et al.*, "A deep neural network accelerator using residue arithmetic in a hybrid optoelectronic system," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 18, no. 4, pp. 1–26, 2022.

[9] H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan, "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2024, pp. 686–703.

[10] G. Yang, C. Demirkiran, Z. E. Kizilates, C. A. R. Ocampo, A. K. Coskun, and A. Joshi, "Processing-in-memory using optically-addressed phase change memory," in *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2023, pp. 1–6.

[11] G. Yang, S. Karimi, C. A. R. Ocampo, A. K. Coskun, and A. Joshi, "Sophie: A scalable recurrent ising machine using optically addressed phase change memory," in *2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2024, pp. 1548–1561.

[12] J. Gu, C. Feng, Z. Zhao, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "Squeezelight: Towards scalable optical neural networks with multi-operand ring resonators," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021, pp. 238–243.

[13] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, "A survey of accelerator architectures for deep neural networks," *Engineering*, vol. 6, no. 3, pp. 264–274, 2020.

[14] 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.

[15] J. Wang, U. Gupta, and A. Sriraman, "Peeling back the carbon curtain: Carbon optimization challenges in cloud computing," in *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, 2023, pp. 1–7.

[16] J. Wang, D. S. Berger, F. Kazhamiaka, C. Irvene, C. Zhang, E. Choukse, K. Frost, R. Fonseca, B. Warrier, C. Bansal *et al.*, "Designing cloud servers for lower carbon," in *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2024, pp. 452–470.

[17] 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.

[18] C. C. Sudarshan, A. Arora, and V. A. Chhabria, "Greenfpga: Evaluating fpgas as environmentally sustainable computing solutions," *arXiv preprint arXiv:2311.12396*, 2023.

[19] L. Eeckhout, "Focal: A first-order carbon model to assess processor sustainability," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2024, pp. 401–415.

[20] P. Dangi, T. K. Bandara, S. Sheikhpour, T. Mitra, and L. Eeckhout, "Sustainable hardware specialization," in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024, pp. 1–9.

[21] M. G. Bardon *et al.*, "Dtco including sustainability: Power-performance-area-cost-environmental score (ppace) analysis for logic technologies," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 41–4.

[22] N. M. Fahrenkopf *et al.*, "The aim photonics mpw: A highly accessible cutting edge technology for rapid prototyping of photonic integrated circuits," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 5, pp. 1–6, 2019.

[23] Y. Zhao, C. Wan *et al.*, "3d-carbon: An analytical carbon modeling tool\\for 3d and 2.5 d integrated circuits," *arXiv preprint arXiv:2307.08060*, 2023.

[24] A. E.-J. Lim, J. Song, Q. Fang, C. Li, X. Tu, N. Duan, K. K. Chen, R. P.-C. Tern, and T.-Y. Liow, "Review of silicon photonics foundry efforts," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 4, pp. 405–416, 2013.

[25] J. A. Cunningham, "The use and evaluation of yield models in integrated circuit manufacturing," *IEEE Transactions on semiconductor manufacturing*, vol. 3, no. 2, pp. 60–71, 1990.

[26] Z. Zhang *et al.*, "Adjoint-based particle defect yield modeling for silicon photonics," in *Optical Modeling and System Alignment*, vol. 11103. SPIE, 2019, pp. 196–205.

[27] S. Y. Siew, B. Li, F. Gao, H. Y. Zheng, W. Zhang, P. Guo, S. W. Xie, A. Song, B. Dong, L. W. Luo *et al.*, "Review of silicon photonics technology and platform development," *Journal of Lightwave Technology*, vol. 39, no. 13, pp. 4374–4389, 2021.

[28] K. Giewont, K. Nummy, F. A. Anderson, J. Ayala, T. Barwicz, Y. Bian, K. K. Dezfulian, D. M. Gill, T. Houghton, S. Hu *et al.*, "300-mm monolithic silicon photonics foundry technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 5, pp. 1–11, 2019.

[29] "Prototyping and low-volume manufacturing of silicon photonic ics," https://www.imeciclink.com/en/asic-fabrication/si.

[30] Z. Yin, M. Zhang, A. Begovic, R. Huang, J. Zhang, and J. Gu, "Simphony: A device-circuit-architecture cross-layer modeling and simulation framework for heterogeneous electronic-photonic ai system," *arXiv preprint arXiv:2411.13715*, 2024.

[31] M. Li, Z. Yu, Y. Zhang, Y. Fu, and Y. Lin, "O-has: Optical hardware accelerator search for boosting both acceleration performance and development speed," in *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 2021, pp. 1–9.

[32] V. Bangari *et al.*, "Digital electronics and analog photonics for convolutional neural networks (deap-cnns)," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–13, 2019.

[33] "Nvidia ampere architecture whitepaper," https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf, 2020.

[34] I. Schneider *et al.*, "Life-cycle emissions of ai hardware: A cradle-to-grave approach and generational trends," *arXiv preprint arXiv:2502.01671*, 2025.

[35] N. P. Jouppi, D. H. Yoon, M. Ashcraft, M. Gottscho, T. B. Jablin, G. Kurian, J. Laudon, S. Li, P. Ma, X. Ma *et al.*, "Ten lessons from three generations shaped google's tpuv4i: Industrial product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 1–14.

[36] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2017, pp. 246–247.

[37] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "Unpu: An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, 2018.

[38] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "Nvidia a100 tensor core gpu: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, 2021.

[39] "Pytorch: Models and pre-trained weights," https://pytorch.org/vision/stable/models.html.

[40] M. O'Connor *et al.*, "Fine-grained dram: Energy-efficient dram for extreme bandwidth systems," in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, 2017, pp. 41–54.