2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)

ECO-CHIP: Estimation of Carbon Footprint of Chiplet-based Architectures for Sustainable VLSI

Chetan Choppali Sudarshan[1], Nikhil Matkar[1], Sarma Vrudhula[1], Sachin S. Sapatnekar[2], and Vidya A. Chhabria[1]
[1]Arizona State University; [2]University of Minnesota

***Abstract*—Decades of progress in energy-efficient and low-power design have successfully reduced the operational carbon footprint in the semiconductor industry. However, this has led to increased embodied emissions, arising from design, manufacturing, and packaging. While existing research has developed tools to analyze embodied carbon for traditional monolithic systems, these tools do not apply to near-mainstream heterogeneous integration (HI) technologies. HI systems offer significant potential for sustainable computing by minimizing carbon emissions through two key strategies: "reducing" computation by "reusing" pre-designed chiplet IP blocks and adopting hierarchical approaches to system design. The reuse of chiplets across multiple designs, even spanning multiple generations of ICs, can substantially reduce carbon emissions throughout the lifespan. This paper introduces ECO-CHIP, a carbon analysis tool designed to assess the potential of HI systems toward sustainable computing by considering scaling, chiplet, and packaging yields, design complexity, and even overheads associated with advanced packaging techniques. Experimental results from ECO-CHIP demonstrate that HI can reduce embodied carbon emissions by up to 30% compared to traditional monolithic systems. ECO-CHIP is integrated with other chiplet simulators and is applied to chiplet disaggregation considering other metrics such as power, area, and cost. ECO-CHIP suggests that HI can pave the way for sustainable computing practices.**

I. Introduction

All aspects of computing, from small chips to large datacenters, come with a carbon footprint (CFP) price tag. For several decades, the semiconductor industry has focused on making chips smaller, faster, and less power-hungry, but few efforts have considered the impact on the environment. The dramatic increase in the demand for compute in the past two decades, fueled by new applications (e.g., artificial intelligence) that demand at-edge and at-cloud-scale computing, has resulted in the information and computing technology (ICT) sector contributing to more than 2% of the world's CFP [1] – half that of the aviation industry [2] and projected to surpass it in the next decade if left unchecked.

Fig. 1 shows the life cycle assessment (LCA) of a semiconductor product and highlights the different sources of greenhouse gases (GHG) in the life of product. The operational costs refer to the CFP generated by the end user, which in the case of a datacenter are the data-to-day activities that draw energy. The embodied costs are the costs that come from design, manufacturing, packaging, and materials sourcing of the server class computation resources in the datacenter.

While technology scaling and electronic design automation have helped to design energy-efficient VLSI systems with

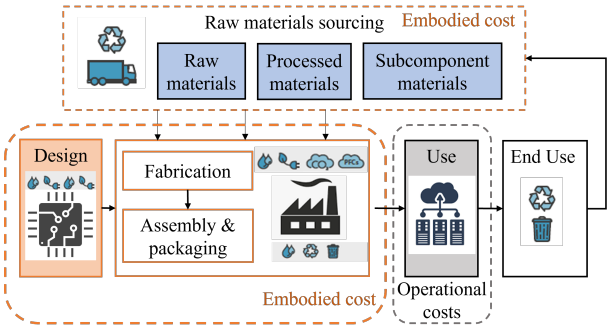

Fig. 1: Embodied and operational CFP sources in the VLSI supply chain [3].

lower operational CFP, the environmental footprint has continued to increase over the past decade and is now dominated by carbon emissions from chip design and manufacturing [3]–[6], i.e., the embodied CFP, especially for low-power edge devices. It is imperative to look beyond the metrics of low power and energy efficiency and include total CFP (embodied and operational) as a first-order optimization metric [4], [7] for sustainable use of today's modern computing devices. Several technology companies have pledged to limit their CFP [6], [8], [9], and this can only be achieved by adopting approaches that are cognizant of CFP.

Further, with Moore's law slowing down and SRAM and analog components not scaling [10], [11], the way forward towards sustaining Moore's law to the era of trillion-transistor multi-functional systems and beyond is through HI [12], [13]. Instead of building system functionality on a single die, HI integrates a set of chiplets, each corresponding to the single die of today, onto a substrate that enables high-density, high-bandwidth chiplet-to-chiplet interconnections. Recent and upcoming advances in HI, including the rapid shrinking of bond pitches between chiplets and interposers [14]–[18], enable the design of increasingly sophisticated integrated systems that not only improve cost due to smaller dies and higher yields but also improve power and performance as shown in recent commercial products [10].

In fact, from a sustainability perspective, heterogeneous chiplet-based systems make a compelling case for CFP evaluation. With the higher yields due to smaller dies, the ability to "mix and match" chiplets in different technology nodes (older nodes have lower defect densities and lower CFP than advanced nodes), lower silicon wastage on the periphery of wafers due to smaller die sizes, and the savings on design costs due to the availability of pre-designed chiplets, HI systems

have the potential to pave a path towards greener VLSI systems. This calls for CFP estimation tools at the architecture level that can model HI **systems** and not just monolithic dies as in [4], [7], [19]. New CFP models are needed to account for packaging overheads, silicon fabrics, and multi-die system integration. Such models can be embedded into the emerging HI design methodologies to optimize HI systems for power, performance, area, cost, *and carbon*.

Inspired by the principles of environmental sustainability – "reduce" and "reuse" – in this paper, we evaluate the potential of HI systems towards sustainable computing. HI systems have the potential to lower CFP by "reducing" carbon emissions by reducing the computation involved in designing each component from scratch and by "reusing" pre-designed and bulk-manufactured general-purpose chiplet blocks through hierarchical approaches. The ability to reuse chiplets across several designs, not only in the current generation of ICs but even in the next generation, can massively amortize the embodied CFP just as it amortizes the dollar cost [20].

In this paper, we introduce ECO-CHIP, a tool to **E**stimate the **C**arb**O**n footprint of **CHIP**let-based architectures for sustainable VLSI. ECO-CHIP is tailored explicitly for heterogeneous systems that incorporate advanced packaging techniques. Our goal is to demonstrate the potential of such systems in reducing CFP compared to large monolithic dies, even after accounting for the overhead associated with packaging. The key contributions of our paper are as follows:

1) To the best of our knowledge, this is the **first work to propose heterogenous chiplet-based systems as a direction toward sustainable VLSI.** The paper highlights how heterogenous chiplet-based architectures enable **"reuse" and "reduced" chip design and manufacturing despite advanced packaging overheads.**
2) We develop a novel architectural-level analysis tool, **ECO-CHIP, to estimate the total CFP (design, manufacturing, packaging, and operation) of HI systems**, accounting for various packaging architectures, scaling, yield, process equipment energy efficiency, lifetimes, duty cycles, wafer silicon wastage, and EDA tool productivity.
3) This is the **first work to build a CFP estimator for a variety of HI packaging architectures,** considering whitespaces on the package substrate/interposer and inter-die communication overheads.
4) We evaluate our tool and the potential HI has in reducing CFP on a diverse set of industry testcases (mobile processors, GPUs, and CPUs) and find that **HI systems can reduce embodied CFP by up to 70%** due to the ability to "mix and match" technology nodes, yield improvements due to smaller die sizes, and availability of predesigned and bulk-manufactured chiplets.
5) We demonstrate the **use of our tool in performing SoC to chiplet disaggregation, considering other metrics such as area, power, and dollar cost** by integrating with other third-party chiplet models.

We open-source ECO-CHIP for broader access and awareness within the community and is currently available at [21].

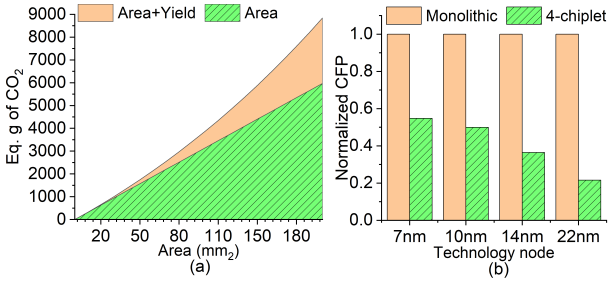

Fig. 2: (a) Embodied CFP versus area of the chip. (b) Comparison of manufacturing CFP of the large monolithic NVIDIA GA102 GPU and a 4-chiplet-based architecture of the GPU.

II. Motivation and Background

A. HI enabling sustainable computing

HI has offered a feasible approach for cost-effective chip design to help sustain Moore's law. A HI system splits a large SoC into multiple smaller dies, referred to as chiplets, where each chiplet may have different functionality, potentially built in different process nodes or designed by separate vendors reducing both design time and cost. All chiplets are integrated into a single package. HI systems have great potential to lower the embodied CFP associated with design and manufacturing when compared to monolithic systems due to several reasons: *(1) Yield and area* As we pack more functionality and logic onto the same monolithic IC, the increase in the area increases the CFP due to an increase in materials needed for manufacturing and a decrease in yield. Fig. 2(a) shows a result for an industry testcase in a 10nm technology (using the techniques described in Section III). We sweep the area of the monolithic SoC up to 200mm^2 and observe an exponential increase in the associated manufacturing CFP (expressed in equivalent grams of CO_2) due to lower yields. In a HI system, each of the smaller dies can be manufactured with a significantly lower environmental cost. For example, Fig. 2(b) compares the CFP of a monolithic NVIDIA GA102 GPU testcase against a 4-chiplet representation of the GPU where the memory and analog components are on independent chiplets, and the large digital block is split into two smaller chiplets. The CFP of the 4-chiplet design is normalized to the monolithic design's CFP for different technology nodes. The 4-chiplet GPU has significantly lower manufacturing CFP even after including the carbon overheads from advanced packaging due to the higher yields when compared to monolithic systems.
(2) Technology node "mix and match" In a chiplet-based system, dies can be implemented in different technology nodes and integrated into a single package. With analog and SRAM blocks not scaling at the same rate as digital logic, several design houses [10], [22] use chiplets in older technology nodes for memory controllers and analog logic. As pointed out in [3], the CFP to manufacture chips in older technology nodes is much lower than for newer technology nodes due to lower defect densities (better yields), fewer lithography steps due to fewer back-end-of-line (BEOL) or front-end-of-line (FEOL) layers, and the better energy-efficiency of lithography equipment involved in manufacturing older technology nodes

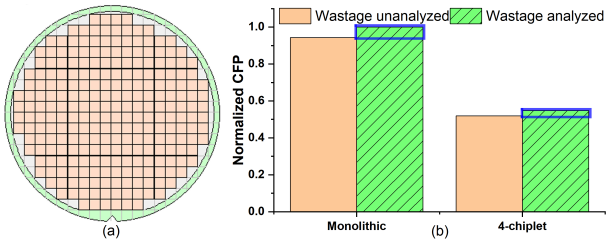

Fig. 3: (a) Dies on a wafer highlighting the green and white regions of the wafer that are wasted. (b) Comparison of the manufacturing CFP with and without analyzing wastage around wafer periphery for GA102 GPU monolithic and 4-chiplet-based architecture on a 450mm wafer.

with today's latest manufacturing equipment. Typically, even EDA tools scale with technology, and the latest versions of EDA tools can perform design faster with better quality of results on an older technology node [23] due to continuous improvements made by the EDA industry.

(3) Chiplet reuse - Reduced design and manufacturing CFP Reusing existing silicon-proven die not only saves design time to market directly but also saves the associated design-time CFP. Moreover, composing custom chips out of small, algorithmic chiplets, reusable across diverse designs, can effectively amortize the non-recurring engineering (NRE) dollar costs and CFP across several different designs [24]. A further reduction in manufacturing CFP is due to the reduction in the area wasted around the periphery of the wafer during manufacturing. Fig. 3(a) shows an image of a wafer with die slices, the green circle around the periphery of the wafer and the white regions are unusable. This area wasted is normalized across the number of dies per wafer (DPW). Smaller-sized dies have lower area wasted compared to larger-sized dies as they can fit a larger number of dies per wafer and also improve the area utilization of the wafer due to the geometric discretization problem. Chiplet-based systems allow a reduction in total number of wafers used due to the fact that more dies can be extracted from a single wafer for smaller die sizes. Fig. 3(b) shows normalized manufacturing CFP of monolithic GA102 testcase and 4-chiplet version of the same testcase with and without considering wastage.

B. HI and packaging architectures

HI systems are available in different packaging architectures, as shown in Fig. 4, varying in cost and complexity. Multi-die HI systems may have anywhere between two to tens of different chiplets. Depending on the number of chiplets, budgeted cost, and complexity, the choice of packaging architecture for heterogeneous systems is different [25]. We describe four advanced packaging and integration technologies:

(1) RDL fanout integration (Fig. 4(a)) involves the integration of multiple chiplets on a package substrate or fanout redistribution layer (RDL) substrate. Typically the package substrate consists of 3-4 RDL metal layers with linewidths and spacings (L/S) varying from 6/6μm to 10/10μm.

(2) Thin film and silicon bridge-based integration uses a package substrate that has thin-film layers defined as embedding fine metal RDLs, or a silicon bridge on top of a build-up organic package substrate or in a fanout epoxy molding compound substrate as highlighted in Fig. 4(b). Intel's embedded multi-die interconnect bridge (EMIB) and TSMC's local silicon interconnect (LSI) are examples of this technology. The package uses local silicon bridges to host ultra-fine L/S structures (2μm) for die-to-die communications.

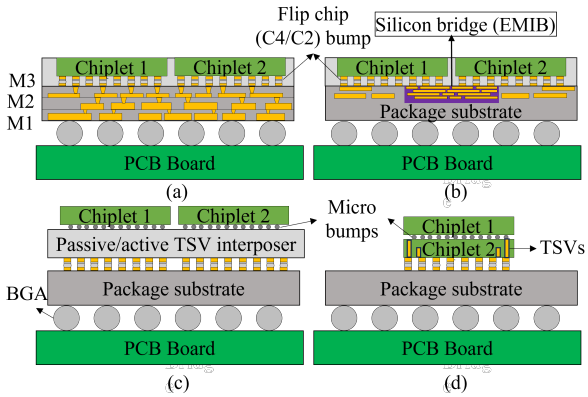

Fig. 4: Packaging architectures: (a) RDL fanout, (b) thin-film and silicon bridge architecture (Intel's EMIB, TSMC's LSI), (c) 2.5D integration with active or passive interposers and (4) 3D stacking with TSVs and microbumps.

(3) Passive and active interposer-based integration involves multiple chiplets in the package that are supported by a through-silicon via (TSV)-less or TSV-based active/passive interposer, and then attached to a package substrate, as shown in Fig. 4(c). This technology is typically termed a 2.5D architecture. The active interposer consists of both FEOL and BEOL layers, while the passive interposer consists of BEOL layers only, both of which are typically implemented in an older technology node.

(4) 3D integration uses active interposers to support the chiplets, which are then attached to the packaging substrate, or stacks multiple chiplets over the packaging substrate, connected through microbumps or through silicon vias (TSVs), as shown in Fig. 4(d), or direct bumpless bonding [16]. With a face-to-back (F2B) stacking of chiplets TSVs are used for connections across tiers; with face-to-face stacking of chips, micobumps are used for inter-die die communication.

C. Summary

HI has opened up a large new design space previously unexplored by architectural-level carbon simulators [4], [7], [19]. This design space theoretically has significant potential to lower CFP. However, the exploration of this space requires the development of models that can account for the different possible design decisions that impact the CFP. For instance, the above four described packaging architectures differ in their yields, assembly process, and material used and therefore have different CFPs. EMIB consists of high-density interconnects with fine L/S, typically having lower yields than the larger RDL layers in fanout packaging. Interposer-based integration strategies typically use more materials, and layers, and have a more complex manufacturing process compared to the fanout RDL and EMIB architectures which result in larger CFP.

Our work, ECO-CHIP, evaluates the CFP for these packaging architectures and the potential HI systems have for

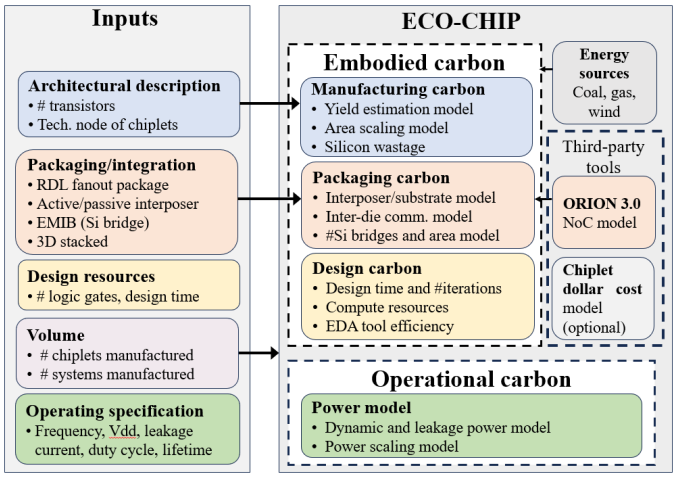

Fig. 5: ECO-CHIP framework highlighting inputs and the models developed to output embodied and operational carbon. The embodied CFP estimation accounts for the CFP from packaging (red), manufacturing (blue), and design (yellow). The operational CFP (green) is estimated from power models.

sustainable computing. ECO-CHIP is integrated with other third-party chiplet-based tools such as ORION 3.0 [26] to model inter-die communication overheads accurately and also integrates with cost [27] and power estimators [26] to perform chiplet disaggregation considering power, area, and cost.

III. ECO-CHIP FRAMEWORK

A high-level description of ECO-CHIP is shown in Fig. 5. It highlights the inputs and the models developed to generate embodied and operational carbon as output. This section describes the ECO-CHIP framework in detail.

A. ECO-CHIP input description

(1) Architectural description High-level description of the SoC or the chiplet-based system, including the predicted number of transistors or logic gates in the chiplet/SoC and the technology node each chiplet is implemented in.

(2) Choice of packaging integration technique ECO-CHIP supports four different advanced packaging architectures (described in Section II). The RDL-fanout-based packaging requires specifying the number of metal layers used and the technology node; the Si bridge-based architectures require specifying the range of the bridge, area of the bridge, and the number of BEOL layers in the bridge; the interposer-based packages require specifying the technology node of the interposer and the number of BEOL layers, and the 3D integration, i.e., chip-to-chip stacking requires specifying the pitch and size of the TSV's and microbumps to estimate the CFP overheads specific to HI.

(3) Operating specification To measure operational carbon, we consider the frequency of operation, lifetime, voltage supply, leakage current, and device usage times as input.

(4) Energy sources and chiplet/system volumes Other inputs include carbon intensity of different sources and volumes of chiplets and systems manufactured to scale the total CFP based on sources of energy and across the number of parts.

B. ECO-CHIP total CFP estimation

With the diversity of today's chips, ranging from low-power processors at the edge where embodied CFP dominates operation CFP to high power consuming GPUs in a datacenter, where operational CFP dominates embodied CFP, optimizing the total CFP (embodied and operational) is essential for generalizable sustainable computing. ECO-CHIP models the total CFP (C_{tot}) as the sum of the operational CFP (C_{op}) and embodied CFP (C_{emb}) as:

$$C_{\text{tot}} = C_{\text{emb}} + \text{lifetime} \times C_{\text{op}} \tag{1}$$

where C_{emb} of the system is the sum of the CFP from the different sources highlighted in Fig. 1 and is given by:

$$C_{\text{emb}} = C_{\text{mfg}} + C_{\text{des}} + C_{\text{HI}} \tag{2}$$

where C_{mfg} is the manufacturing CFP of all chiplets, C_{des} is the design CFP of all chiplets, and C_{HI} is the overhead from HI including contributions from manufacturing and assembly of the advanced package and any inter-die communication.

The operational CFP, C_{op}, is modeled as the carbon intensity of the source of energy during usage ($C_{\text{src,use}}$) times the energy spent during usage (E_{use}) and is given by:

$$C_{\text{op}} = C_{\text{src, use}} \times E_{\text{use}} \tag{3}$$

In the rest of this section, we detail the models for each of the components in Eqs. (1), (2), and (3).

C. ECO-CHIP manufacturing CFP estimation

We model the manufacturing CFP of the heterogeneous system as the sum of the manufacturing CFP of each chiplet, i, and is given by $C_{\text{mfg}} = \sum_{i=1}^{N_C} C_{\text{mfg},i}$. To estimate the manufacturing CFP of each chiplet, we make three essential modifications to [4], [7] to support the estimation of embodied CFP and perform disaggregation as described below:

(1) Area scaling models Since a system disaggregation algorithm or a heterogeneous system requires selecting a technology node for each chiplet, our carbon estimation tool uses transistor density scaling trends from [28], [29] and transistor counts from our testcase architectures to determine the area of a chiplet in a specific technology node. The area scaling models are critical to the estimation of CFP as larger chiplet areas in older technology nodes can have larger CFP even though they have lower CFP per unit area (CFPA). We use three different area scaling models for logic, memory, and analog blocks, as each has different transistor densities and, therefore different areas with every technology node. We evaluate the area of the die as $A_{die}(d,p) = D_T(d,p) \times N_T$, where $D_T(d,p)$ is the transistor density for design type d and process p, N_T is the number of transistors in the die, and $A_{die}(d,p)$ is the area of die of type d in process node p.

(2) Yield models One of the primary advantages of HI is the cost savings that come with larger manufacturing yields due to smaller die sizes. The increase in yield compared to a large monolithic die also helps lower CFP. However, if the die is in an older technology node, then $A_{\text{die}}(d,p)$ must be accounted for as an increase in the area may lower yields which also lowers the CFP, as shown in Fig. 2(a). To estimate the impact

of the area on yield and CFP, we use a negative binomial yield distribution model given by [30]–[32]:

$$Y(d,p) = \left(1 + \frac{A_{\text{die}}(d,p) \times D_0(p)}{\alpha}\right)^{-\alpha} \quad (4)$$

where $Y(d,p)$ is the yield of die with area $A_{die}(d,p)$, $D_0(p)$ is the defect density for process p, α is a clustering parameter. It is important to note that the defect density depends on p and scales with technology. This dependence is important to capture because legacy nodes have lower defect densities which result in larger yields, but older technology nodes result in larger A_{die} values leading to lower yields. ECO-CHIP considers these tradeoffs while estimating CFP.

(3) Energy-efficiency of process equipment With advances in process equipment, the energy efficiency of the photolithography equipment improves at every step, especially for the more mature technology nodes. We incorporate the energy efficiency of the equipment as a derate factor (η_{eq}) from [33]. The $C_{\text{mfg},i}$ on a per chiplet basis is given by the sum of the product of the carbon footprint per unit area (CFPA) of manufacturing a die and the area of the die and the product of the CFPA of silicon (CFPA_{Si}) and amortized area of silicon wafer wasted (A_{wasted}):

$$C_{\text{mfg},i} = \text{CFPA} \times A_{\text{die}}(d,p) + \text{CFPA}_{\text{Si}} \times A_{\text{wasted}} \quad (5)$$

$$\text{CFPA} = \frac{(\eta_{eq} \times C_{\text{mfg, src}} \times \text{EPA}(p) + C_{\text{gas}} + C_{\text{material}})}{Y(d,p)} \quad (6)$$

where $C_{\text{mfg, src}}$ is carbon intensity which depends on the energy source of the fab (i.e., renewables vs. non-renewables), which converts the energy consumed into carbon emission. EPA is the energy consumed per unit area during manufacturing of process p and derived from [5], C_{gas} is the CFP from the greenhouse gas emissions, and C_{material} is the carbon footprint of sourcing the materials for fabricating the chip/chiplet.

The wasted silicon area in a wafer of area A_{wafer} is highlighted in Fig. 3(a) by the green and white regions. The die cannot occupy zones within its half die diagonal, reducing the usable diameter by $L_{\text{d}}/\sqrt{2}$. Therefore, the number of dies per wafer (DPW) and the amortized area wasted per die (A_{wasted}):

$$\text{DPW} = \left\lfloor \frac{\pi(\frac{D_{\text{wafer}}}{2} - \frac{L_d}{\sqrt{2}})^2}{A_{\text{die}}} \right\rfloor \quad (7)$$

$$A_{\text{wasted}} = \frac{A_{\text{wafer}} - (\text{DPW} \times A_{\text{die}})}{\text{DPW}} \quad (8)$$

where D_{wafer} is diameter of the wafer, L_d is the side length of the die, and A_{die} is the die area. An important observation here is that smaller dies have lesser A_{wasted} compared to larger dies as we can cramp in more DPW. This allows for a larger amortization of the wasted.

D. ECO-CHIP HI-oriented CFP overheads

With the widespread adoption of HI systems, the cost of packaging is projected to dominate design [34]. Although there are several sustainability reports from large semiconductor manufacturing and design companies, these reports do not specifically break down the contributions from packaging. The prior art in this area has been limited to wire bond packages and flip chip packages [35]. Since HI has opened up a previously unexplored design space, it requires developing models that can account for the different possible design decisions in the HI system that impact the CFP. In particular, decisions related to the choice of the package (C_{package}), whitespace on the package substrate or interposer ($C_{\text{whitespace}}$), and inter-die communication ($C_{\text{mfg, comm}}$). In our work, we measure the CFP from these three sources as described below:

(1) Package-related overheads (C_{package}) We develop models for the four different packaging architectures, described in Section II based on architectural descriptions, materials, and packaging technology nodes from [25], CFP estimates from [5], and packaging industry reports [36]–[38].

(a) RDL Fanout: This packaging architecture uses an epoxy molding compound (EMC) substrate with RDL metal layers patterned to make connections between the chiplets as shown in Fig. 4(a). Our CFP model uses the energy per unit area per metal layer (EPLA) from a manufacturing fab to determine CFP overheads with the RDL layers. Based on the number of layers, the yield of the layers, and EPLA, we determine the embodied CFP of an RDL package as:

$$C_{\text{RDL}} = \frac{L_{\text{RDL}} \times \text{EPLA}_{\text{RDL}}(p) \times C_{\text{pkg, src}} \times A_{\text{package}}}{Y(\text{RDL},p)} \quad (9)$$

where $\text{EPLA}_{\text{RDL}}(p)$ is the energy consumed in patterning a single RDL layer in process p per unit area, $C_{\text{pkg,src}}$ is the carbon intensity of the packaging fab which is based on the source of energy (renewable or non-renewable sources), L_{RDL} is the number of layers of RDL in the package substrate, $Y(\text{RDL},p)$ is the yield of the RDL in process p estimated using Equation (4), and A_{package}. The area of the package substrate is estimated after considering the whitespace and routing overheads and described later in this section.

(b) Silicon bridge: A silicon bridge is a high-density interconnect between two chiplets, and we model its CFP similar to the CFP of the RDL fanout-based package except that they have lower linewidth and spacing (L/S) and, therefore, lower yields when compared to RDL fanout. Our model uses the EPLA values from [5] for an advanced technology node lower metal layer with ultra-fine L/S. These high-density interconnects do not span the entire area of the package substrate but are local to a region in the package based on the floorplan of the chiplets. The number of silicon bridges and their placement depends on the chiplet floorplan and bandwidth requirements. In our work, we consider bridge ranges and typical bridge areas from Intel's EMIB silicon bridge specification [39] as input to determine the number of bridges that must be used. An additional bridge is added if the two adjacent dies that must be connected through silicon bridges have overlapping die edges larger than the range. The CFP of a silicon bridge-based packaging architecture is given by:

$$C_{\text{bridge}} = \frac{N_{\text{bridge}} \times L_{\text{bridge}} \times \text{EPLA}_{\text{bridge}}(p) \times C_{\text{pkg, src}} \times A_{\text{bridge}}}{Y(\text{bridge},p)} \quad (10)$$

where L_{bridge} is the number of metal layers in the bridge, A_{bridge} is the area occupied by the silicon bridge in the package substrate, N_{bridge} is the number of silicon bridges, $Y(\text{bridge}, p)$ is the yield of fabricating the silicon bridge in process p in the bridge cavity, $\text{EPLA}_{\text{bridge}}(p)$ is the energy per unit layer per unit area of patterning the silicon bridge in process p.

(c) Active interposer: Active interposers are manufactured to include transistor devices within the interposer, providing several unique capabilities not possible with passive interposers. We model these interposers as an additional large die that is typically in an older technology node compared to the chiplets. However, unlike a regular chiplet, the active region is only restricted to local regions with routers and repeaters.

We use a similar model based on Eq. (6) to estimate CFP overhead from active interposer. Interposer-based packaging architectures have higher CFP when compared to the RDL fanout-based packaging and EMIB-based packaging as the interposer acts as an additional large silicon die that spans the entire area of all the chiplets put together with BEOL layers across the entire interposer and active FEOL layers locally in those areas that have routers or repeaters.

(d) Passive interposer: Unlike active interposers, passive interposers only contain metal interconnect, so they cannot include active logic like routers, or repeaters in the interposer. We model the CFP of the passive interposer in a similar way as Equation (9) on a per unit area and per layer basis.

(e) 3D integration: This packaging architecture stack chiplets as shown in Fig. 4(d) to minimize the 2D footprint and maximizes bandwidth where inter-chiplet communication is performed with TSVs, microbumps, or hybrid bonds. Our CFP model uses the energy per unit area per metal layer (EPLA) from a manufacturing fab to determine CFP overheads with the TSVs, microbumps, and hybrid bonds [5], [7]. The number of TSVs, bumps, or bonds depends on the size of the chip and its pitches. In our work, we assume a dense network of TSVs, bumps, or bonds placed at the minimum pitch [18] to maximize inter-chiplet bandwidth. We use the TSV diameter and pitch values [18], [40], hybrid and microbump pitch values from [41] and EPLA to determine the embodied CFP as:

$$C_{\text{3D}} = \frac{N_{\text{TSV, bump, bond}} \times \text{EPA}_{\text{TSV, bump, bond}}(p) \times C_{\text{pkg, src}}}{Y(\text{3D}, p)} \tag{11}$$

where $N_{\text{TSV, bump, bond}}$ is the number of TSVs/bumps/bonds, $Y(\text{3D}, p)$ is the yield of the 3D package assembly accounting for misalignments of bumps the bonding yield between chiplets. $\text{EPA}_{\text{TSV, bump, bond}}(p)$ is the energy per unit area of patterning the TSV or manufacturing the bump in process p.

(2) Inter-die communication overheads ($C_{\text{mfg, comm}}$) Unlike EMIB and RDL-based packaging architectures, which are limited to supporting few (four - five) chiplets [25] interposer based 2.5D architectures support tens of chiplets while 3D integration techniques can have 2-3 tiers of logic. However, both interposer-based techniques and 3D integration techniques come with large inter-die communication overheads which require are protocols such as network-on-chip (NoC). To support an NoC router, each chiplet must be equipped with a network interface controller (NIC). In passive interposers, router modules must be placed within the chiplets, contributing to chiplet area and degrading yield and $C_{\text{mfg},i}$ while with active interposers, router modules can be moved from the chiplets to the interposer, reducing the area in the chiplets and therefore improving chiplet yield and $C_{\text{mfg},i}$ compared to passive interposers.

To estimate the CFP overheads of routing, we use a third-party tool, ORION 3.0 [26] and [42]. ORION is used to estimate the power overhead due to the additional inter-die communication NoC circuitry and [42] is used to estimate the area overhead. The work [42] models the network on interposers (NoI/NoC) area by including flit width, bidirectional port counts, and microbump pitches. However, it provides area values for only a small set of specific technology nodes (11nm to 65nm), we scale the values for the technology nodes we consider. The NoC area is added to either the chiplets or the interposer based on active or passive interposer and implemented in the same technology node as that of the chiplet or the interposer. Although ORION 3.0, supports 45nm and 65nm technologies, we modify the parameter files for the appropriate technology node. ORION 3.0 models the power of the NoC by estimating the number of instances based on the microarchitectural parameters, including the number of ports, flit width, and buffers [26], [42], [43]. The CFP overhead for interposer-based NoC routers for inter-die communication is given by: $C_{\text{mfg, comm}} = \text{CFPA} \times A_{\text{router}}(d, p)$, where CFPA is defined in (6). For the passive interposer, $A_{\text{router}}(d, p)$ is added to the area of the chiplet, after which yield and $C_{\text{mfg, i}}$ is calculated. For active interposers, the carbon contribution of $A_{\text{router}}(d, p)$ is used to add to the embodied CFP of the system.

It's important to note that for passive interposers, the NoC is implemented in the same technology node as the chiplet, which is a more advanced node than those routers that are a part of the package. Therefore, routers for passive interposers are of lower areas than the active interposer router in an older technology node. For EMIB- and RDL-based packages, there are additional communication overheads for PHY [39] interfaces that are typically part of the chiplet itself. These interfaces are typically designed as IPs and have small additional areas when compared to the chiplets.

(3) Whitespace overheads ($C_{\text{whitespace}}$) To estimate the area of the package substrate or interposer, ECO-CHIP uses a whitespace or system area estimation algorithm. The algorithm performs recursive bi-partitioning to build a slicing floorplan [44] of the chiplets on the package substrate/interposer. An initial two-way partition is created by assigning the chiplets (sorted in decreasing order of their area), one by one, to the partition with a lesser total weight. Our model uses the area of each partition as the weight, which results in an area-balanced initial partition. The bi-partitioning procedure is then used recursively within each partition to perform a K-way partition of the chiplets by first creating two equal-sized partitions, then independently dividing each of these into two subpartitions each, and so on till a partition contains only one chiplet. This effectively creates a full binary tree where each leaf node is

a chiplet and each internal node represents a partition. The overall floorplan and its area can be derived by processing the partitions and chiplets within the tree.

For each leaf node, processing involves setting the orientation and aspect ratio of the chiplet to get a bounding box. At the internal nodes, this involved combining two subpartitions together, accounting for whitespace overheads. There are two sources of whitespace overheads: (i) spacing between two subpartitions due to chiplet spacing constraints [42], [45], (ii) if the two subpartitions are imbalanced in terms of their dimensions, we create a bounding box of the two partitions which will result in additional whitespace. The recursive bipartisan floorplan also provides us with interfaces between each pair of chiplets to identify locations for routers, and silicon bridges on the package substrate/interposer.

E. ECO-CHIP design CFP estimation

Although design CFP (C_{des}) is amortized across the number of chiplets of type i manufactured (N_{M_i}) and systems manufactured (N_S), several cutting-edge accelerators, GPUs, and server CPUs are not manufactured in sufficiently large numbers to amortize the cost of design across the number of parts manufactured. We model the design CFP, C_{des} as:

$$C_{\text{des}} = \sum_{i=1}^{N_C} \frac{C_{\text{des},i}}{N_{M_i}} + \frac{C_{\text{des,comm}}}{N_S} \tag{12}$$

where $C_{\text{des},i} = t_{\text{des},i} \times P_{\text{des}} \times C_{\text{des,src}}$ is the design CFP of a single chiplet, $C_{\text{des,comm}}$ is the design CFP of routers for inter-die communication, $t_{\text{des,i}}$ is the CPU compute time it takes to design a chip/chiplet, P_{des} is the power consumed by the compute resources (CPUs) used to design the chips, $C_{\text{des,src}}$ is the CFP of the energy source. We model $t_{\text{des, i}}$ as:

$$t_{\text{des,i}} = \frac{t_{\text{verif},i} + (t_{\text{SP\&R,i}} + t_{\text{analyze},i}) \times N_{\text{des}}}{\eta_{\text{EDA}}} \tag{13}$$

where $t_{\text{verif},i}$, $t_{\text{SP\&R},i}$, $t_{\text{analyze},i}$ are the compute time for verification, and a single synthesis, place, and route (SP&R) run and a single simulation of all analysis respectively, and N_{des} is the number of design iterations. Further, to model the EDA tool improvement with new version releases, we create a near-linear regression model based on productivity for different technology nodes [23] and scale the value of $t_{\text{des,i}}$ by η_{EDA}.

F. ECO-CHIP operational CFP estimation

We estimate the operational CFP (C_{op}) by modeling for the total energy of the system during usage E_{use} based on the operational specification give by:

$$E_{\text{use}} = T_{ON} \times (V_{dd} I_{\text{leak}} + \alpha C V_{dd}^2 f) \tag{14}$$

where T_{ON} is the time for which the system is ON, V_{dd} is the supply voltage, f is the average use case frequency of operation of the system (since most systems are not operating at maximum frequency throughout their use), α is average switching activity, C is the load capacitance, and I_{leak} is the leakage current of the system. We then substitute E_{use} in Eq. (3) to estimate the operational CFP. E_{use} also includes the HI-related power overheads, such as inter-die communication modules (NoCs). For battery-operated devices such as a mobile processor, we can directly estimate E_{use}, based on the battery rating and frequency of recharging [4], [7].

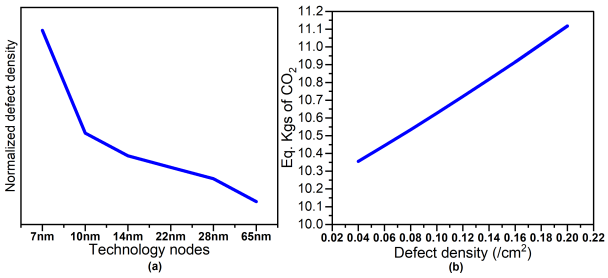

Fig. 6: (a) Normalized defect density with technology node scaling. Older nodes have lower defect densities [31], [32]. (b) Impact of defect density on total CFP.

IV. ECO-CHIP SETUP AND REAL-WORLD TESTCASES

(1) Input parameters ECO-CHIP uses several input parameters which are listed in Table I with their supported range of values and their sources. For instance, based on the source of energy, whether it is coal, gas, wind, etc., the $C_{\text{mfg, src}}$ can be a value between 30g – 700g of CO_2 or based on the technology node, the defect densities can be between 0.07 – 0.3/cm^2 [31], [32]. Defect density varies across nodes, and they tend to mature over time. Fig. 6(a) shows the normalized defect density trend over different technology nodes. For a given technology node, defect densities begin to plateau as the node matures with time [46]–[48]. Fig. 6(b) shows the variation in total CFP as a function of defect density. Although our simulator can handle a range of technology nodes for packaging and a range of derate factors for $C_{\text{mfg, src}}$, our results in this section are shown for specific values, i.e., we assume all packaging technology (RDL, EMIB, and active/passive interposers) to be in 65nm, the D_{wafer} to be 450mm, and the energy source is from coal at 700g of CO_2 per KWh. Based on the testcase, we vary the technology node for each of the chiplets to explore the possible design space and estimate $C_{\text{mfg},i}$. Based on the technology each chiplet is implemented in, we choose the appropriate values from the specified ranges.

(2) Testcases and architectures We evaluate our carbon simulator on four industry testcases: (i) Intel server-class 2-chiplet-based CPU, Emerald Rapids (EMR) [52] (to be released in Q4 2023), (ii) NVIDIA GA102 GPU (2020) [53], (iii) Apple A15 SoC (2021) [54] and (iv) AR/VR (2022) [55]. The input to our simulator is an architectural description of these testcases with the die area breakdowns for each of these processors. We obtain the area breakdowns of each of these testcases from third-party websites such as [52], [53], [56], [57].

For the monolithic SoCs (GA102, and A15) we break them into chiplets based on the block-level architecture. We use one chiplet for memory, another chiplet for analog components, and a third chiplet for digital logic inspired by [10]. For our 3-chiplet testcases we follow a three-tuple convention such as (7, 10, 14), which indicates the technology nodes the (digital, memory, analog components) are implemented in, respectively.

TABLE I: Input parameters to ECO-CHIP and their range of values.

Model	Parameter	Value	Unit	Source
$C_{\text{mfg},i}$	$D0(p)$	0.07 – 0.3	/cm^2	[31], [32]
	α	3		[31], [32]
	$D_T(d,p)$	5 – 150	MTr/mm^2	[28], [29]
	$\eta_{eq}(p)$	0 – 1		[33]
	$C_{\text{mfg, src}}$	30 – 700	g CO$_2$/kWh	[4], [5]
	EPA(p)	0.8 – 3.5	kWh/cm^2	[4], [5]
	C_{gas}	0.1 – 0.5	kg CO$_2$/cm^2	[4], [5]
	C_{material}	0.5	kg CO$_2$/cm^2	[4], [5]
	D_{wafer}	25-450	mm	[49]
C_{package}	RDL tech.	22nm - 65nm		[25], [39], [42]
	EPLA$_{\text{RDL}}(p)$	0.05 – 0.2	kWh/cm^2	[4], [5]
	$C_{\text{pkg, src}}$	30 – 700	g CO$_2$/kWh	[4], [5]
	L_{RDL}	3 – 9		[25]
	L_{bridge}	3 – 4		[39]
	Bridge tech.	22nm – 65nm		[39]
	EPLA$_{\text{bridge}}(p)$	0.1 – 0.35	kWh/cm^2	[4], [5]
	Bridge range	2×2	mm^2	[39]
	TSV pitch	10 – 45	μm	[18], [40]
	Microbump pitch	10 – 45	μm	[18]
	Hybrid bond pitch	1 – 10	μm	[41]
$C_{\text{mfg,comm}}$	Interposer tech.	22nm – 65nm		[42]
	NoC flit width.	512 bits		[42]
$C_{\text{whitespace}}$	Chiplet spacing	0.1 – 1	mm	[42], [45]
C_{des}	η_{EDA}	0 – 1		[23]
	P_{des}	10	W	[50]
	N_{des}	100		[51]
	$C_{\text{des, src}}$	30 – 700	g CO$_2$/kWh	[4], [5]
$C_{\text{operational}}$	V_{dd}	0.7 – 1.8	V	
	T_{ON}	5% - 20%		
	Lifetime	2 – 5	years	

For EMR, an EMIB-based 2-chiplet testcase, we perform CFP estimation on the original architecture as is.

V. Evaluation of potential CFP savings due to HI

We evaluate total CFP and highlight the new design space and CFP savings chiplet-based technologies enable through technology node mix and match, different choices of packaging architecture, and chiplet reusability.

A. Chiplet technology space exploration for reduced CFP

We demonstrate how the ability to mix and match technology nodes for different chiplets in a system improves embodied CFP. At the same time, we will compare ECO-CHIP embodied CFP with the existing embodied CFP estimator, ACT [4], [7], and highlight how ACT grossly miscalculates the CFP as it does not model for package assembly, wafer area wasted, and design CFP. As an example case study, we use a 3-chiplet version of NVIDIA GPU GA102 with RDL fanout-based packaging architecture to evaluate the various components of CFP for various chiplet disaggregation scenarios.

(1) Manufacturing and HI-related CFP The manufacturing (chip and package) CFP of GA102 with RDL fanout packaging, for different configurations of technology nodes for each chiplet, is shown in Fig. 7(a). The x-axis lists the three-tuple configuration listing the technology node each chiplet is implemented in. The (7,7,7) scenario is a monolithic representation of the architecture of a single die in a 7nm node. It, therefore, does not have the additional HI-related packaging overheads. The figure shows that the lowest C_{emb} is for the (7, 14, 10) scenario. This is because the analog components and memory blocks [10] do not scale in the area as much as the digital blocks and can therefore be implemented in an older technology node with almost the same area. On the contrary, in the (10, 10, 10) scenario, the digital logic scales to a much larger area and therefore has a larger CFP than even the monolith resulting in a larger CFP.

From this result, it is clear that HI enables using chiplets that have smaller areas and higher yields, which helps lowers the CFP, and the further integration of chiplets in different technology nodes can further lower the CFP as older nodes have lower EPA than advanced nodes. ACT [4], [7], uses a fixed value of package assembly CFP (150g of CO$_2$) irrespective of the area of the package, or type of packaging architecture, or the wafer area wasted and therefore can inaccurately estimate C_{mfg} by at least 10kg of CO$_2$ emission ($\approx 20\%$ of C_{emb}).

(2) Design CFP From our experiments in performing SP&R of large designs, we find that the $t_{\text{SP\&R},i}$ for a design with 700,000 logic gates in a 7nm commercial technology is about 24 CPU hours. These estimates are on a 192GB RAM machine with a dual-core Intel Xeon CPU with 8 threads, each running at a 2.4GHz clock frequency. Therefore, extending this model to the GA102 testcase, $t_{\text{SP\&R},i} = 1.5 \times 10^5$ CPU hours as it has over 4.5B logic gates. Assuming $P_{\text{des}} = 10$W [50] and the energy supplied comes from non-renewable sources, then a single run of SP&R results in 8,400kg of CO$_2$ equivalent emission in the 7nm technology node. Fig. 7(b), shows the design carbon for a **single iteration** of SP&R for the 3-chiplet testcase. Older technology nodes have lower design times due to EDA tool scaling [23], and therefore, have lower CFP compared to the monolithic SoC in an advanced 7nm technology. In addition, since HI enables the "reuse" of predesigned chiplets, in principle, the same chiplet can be reused for another design saving the entire associated C_{des}.

Although the C_{des} values in Fig. 7(b) are significantly large, these costs are amortized across the number of parts manufactured (N_S). The figure only shows the results for a single iteration of SP&R. However, with hundreds ($N_{\text{des}} = 100$) of design iterations and SP&R runs and verification dominating 80% of the product development time, the design of an IC can easily contribute to over 2,000,000kg of CO$_2$ equivalent emission, assuming all compute energy is coming from non-renewable sources. Assuming the number of manufactured parts is $N_S = 100,000$, the SP&R carbon cost gets amortized to 12kg of CO$_2$ equivalent emission per IC, which is more than 25% of C_{mfg} (see Fig. 7(a)). This significant contributor to C_{tot} has not been considered in ACT [4], [7].

(3) Embodied CFP To estimate C_{emb}, we sum the C_{mfg} and C_{HI} CFP from Fig. 7(a) and amortized C_{des} assuming, $N_p = 100,000$ and $N_{\text{des}} = 100$ from Fig. 7(b). Fig. 7(c) shows the C_{emb} for different configurations of the 3-chiplet GA102 testcase. C_{emb} is compared against ACT [4], [7] C_{emb}. Since ACT does not estimate C_{des} and packaging-related CFP, it inaccurately estimates a lower CFP.

(4) Total CFP Fig. 7(d) shows C_{tot} split into C_{op} and C_{emb} components. Given the power-hungry GPU [58], with a maximum power rating of 450W and average $E_{use} = 228$kWhr, the embodied carbon is approximately 20% of C_{tot}. HI lowers the C_{emb} compared to a monolithic SoC but increases the C_{op} due to communication overheads and use of chiplets implemented in old technology nodes (larger supply voltages). For the

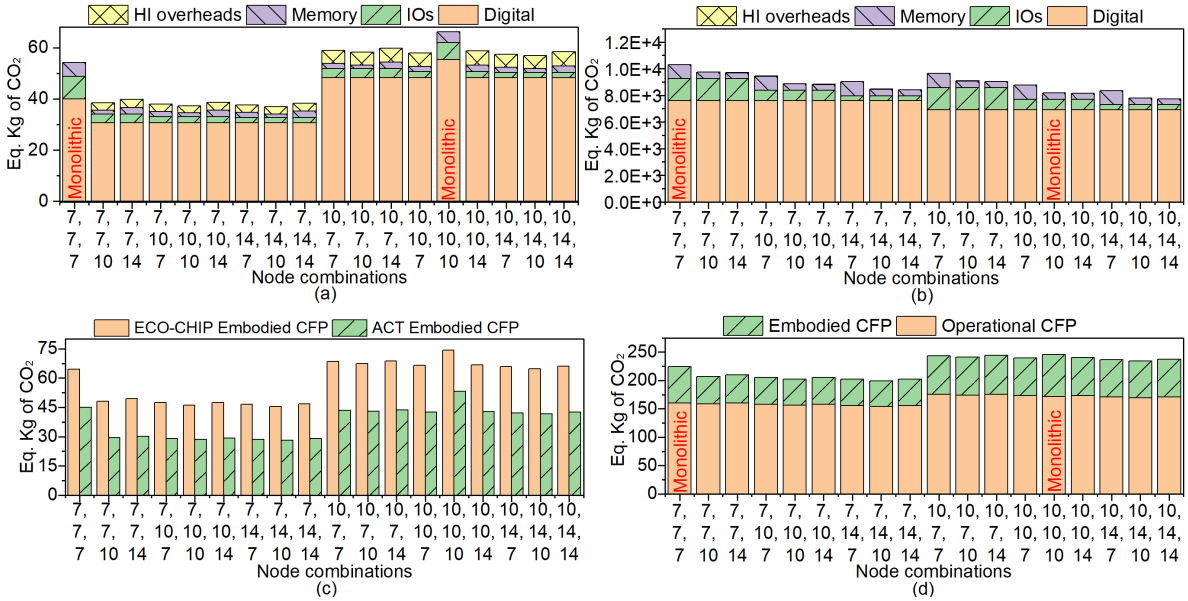

Fig. 7: (a) C_{mfg} and C_{HI}, (b) C_{des} for a **single iteration** of SP&R, (c) C_{emb} for different configurations of three chiplets (C_{des} uses $N_{\text{iter}} = 100$, and $N_s = 100,000$) compared with C_{emb} from ACT [4], [7], and (d) C_{tot} split into its C_{op} and C_{emb} for the GA102 3-chiplet architecture with RDL fanout.

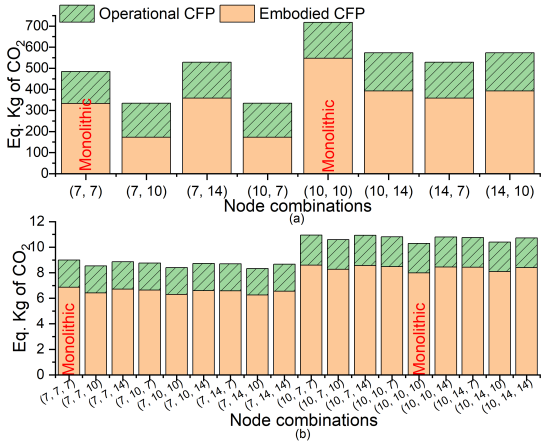

Fig. 8: Total CFP compared to monolithic counterparts for (a) EMR 2-chiplet with EMIB packaging (b) A15 mobile processor with RDL fanout packaging.

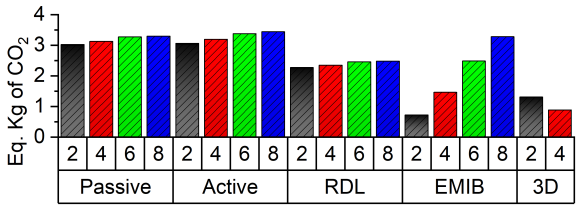

Fig. 9: HI-related CFP overheads for different N_c values.

GA102 testcase, the decrease in C_{emb} dominates the increase in C_{op} (over a two-year lifetime), making the HI system more sustainable than the monolith.

In low-power battery-operated devices, C_{emb} dominates C_{op} [4], [6], and savings in the C_{emb} significantly lower C_{tot}. For example, Fig. 8, shows C_{tot} split into C_{op} and C_{emb} for (a) the 2-chiplet EMR testcase and (b) a 3-chiplet version of A15 mobile processor, both compared to their monolithic counterparts. For C_{des}, we assume $N_S = 100,000$ and $N_{\text{iter}} = 100$. The E_{use} value is obtained from battery specification and a battery charging rate for the mobile phone, and for the EMR testcase it is obtained by profiling a server-class CPU. The figure shows that the mobile processor has a lower operational footprint percentage (40%), unlike the CPU/GPU processors. The improvements in C_{emb} due to technology mix and match are lower in A15 compared to GA102, as it is smaller in area.

(5) Key takeaways (a) We find that design and package assembly CFP are significant components to total CFP and cannot be ignored as in [4], [7]. (b) Chiplets implemented in different technology nodes, lower EPA, improve yield, and provide a whole new design space to explore. (c) Similar to the insight in [27] concerning dollar cost, we find that larger SoCs are more suited to benefit from C_{emb} savings when disaggregated into chiplets when compared to smaller SoCs. The C_{emb} of GA102 lowers by 30% when compared to its monolithic counterpart. (d) Low-power SoCs have a lower C_{op} to C_{emb} ratio and are more suited to benefit from a reduction in total CFP when disaggregated to chiplets.

B. Packaging technology space exploration for reduced CFP

(1) HI-related CFP overheads for different packaging types Although the choice of packaging architecture is driven by application requirements such as bandwidth, area, and power, the CFP for different packages varies significantly and can be considered a metric to drive early architectural decisions. To understand the differences in CFP overheads of the five packaging architectures considered, we use the large digital logic component of GA102 as an example testcase. We split the 500mm^2 monolithic digital logic block into N_c different chiplets and evaluate C_{HI}. Fig. 9 shows the difference in CFP for these architectures separated by routing overheads and package-related overheads (whitespace and area). All the interconnects in the package substrate are modeled in a 65nm technology for the five packaging architectures.

Silicon-bridge-based (EMIB-based) architectures have the least CFP for 2-chiplet-based architectures of the 500mm^2

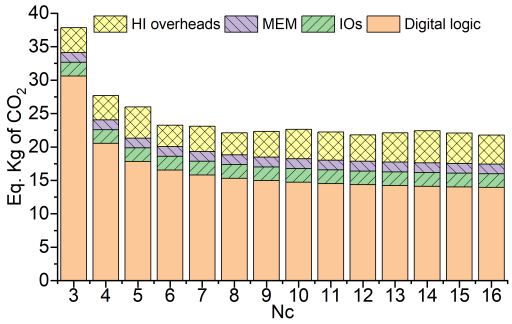

Fig. 10: Manufacturing CFP (C_{mfg} and C_{HI}) for N_c chiplets in GA102.

monolith testcase. However, as N_c increases, the number of silicon bridges also increases, and CFP increases. The RDL-based packages have the least overheads for the 6- and 8-chiplet architectures, but due to their architecture definition, they have lower communication bandwidth when compared to silicon bridges or interposers. Therefore, based on the bandwidth requirements of the testcase, such tradeoffs between performance and CFP can be considered using ECO-CHIP. The figure also shows that the passive interposer has lower routing overheads as the router is part of the chiplet and is in the same technology node of the chiplet. Therefore, in passive interposer technologies, due to the advanced node (7nm in this testcase) in which the router is implemented, the area overheads are smaller than the 65nm router in the active interposer. The routing overheads of RDL, passive interposer, and silicon-bridge (EMIB) architectures are small and near-negligible compared to the core chiplet areas. For the 3D package, we sweep the number of tiers/chiplets (from two to four); as N_C increases, the 2D area of each chiplet reduces, and more chiplets are stacked to implement the same logic. Therefore, the CFP decreases despite the reduction in overall package yield due to an increase in the number of TSV/bumps (the package yield is the product of the yield of each tier).

(2) Manufacturing CFP and HI overheads with N_c chiplets In addition to the 3-chiplet architecture of GA102, we also evaluate the C_{mfg} and C_{HI} for $N_c > 3$ where the digital logic block is further split it into smaller chiplets each implemented in 7nm. The analog (IOs) and memory chiplets are in 14nm and 10nm, respectively. Fig. 10 shows C_{mfg} and C_{HI} for different N_c. As N_c increases, the $C_{\text{mfg},i}$ decreases due to smaller chiplets and better yields, while C_{HI} overheads increase. The data shows that beyond a certain chiplet size (or N_C), the CFP savings reduces as C_{HI} dominates.

(3) Packaging technology parameters and its impact on CFP C_{HI} for the different packaging architectures supported by ECO-CHIP are based on the estimated area overhead and the computed package yield. For each packaging architecture, certain key parameters determine their assembly CFP. For instance, L_{RDL} directly affects C_{RDL} (Eq. (9)) as shown in Fig. 11(a). The figure sweeps the number of BEOL layers in the RDL fanout package from 4 to 9, showing the linear increase in C_{HI}. Fig. 11(b) shows the decrease in C_{HI} with an increase in EMIB range/pitch. The increase in range reduces the number of bridges needed for inter-die communication lowering C_{HI}. Fig. 11(c) shows the difference in HI-related

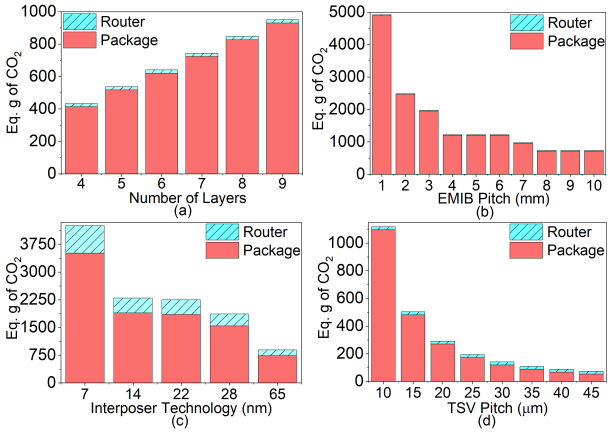

Fig. 11: C_{HI} for the A15 testcase with different parameter sweeps. C_{HI} for different: (a) L_{RDL} for RDL-fanout, (b)bridge ranges for EMIB, (c) interposer technology nodes for active interposer, and (d) TSV pitches for 3D.

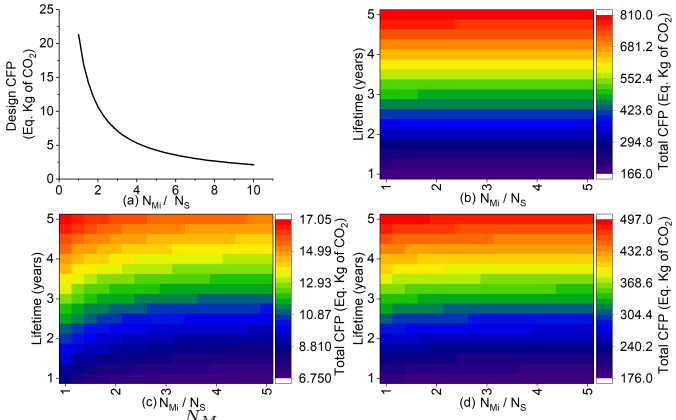

Fig. 12: (a) C_{des} vs. $\frac{N_{M_i}}{N_S}$ highlighting the reduction in C_{des} with the increase in manufacturing volume. Variation in C_{tot} as a function of $\frac{N_{M_i}}{N_S}$ and lifetime for (b) GA102, (c) A15, and (d) EMR 2-chiplet testcases in 7nm.

overheads for active interposers implemented in different technology nodes. Older technology nodes have lower EPA and therefore, lower CFP. Fig. 11(d) sweeps the TSV pitch. Larger TSV pitches imply fewer TSVs between the two tiers ($N_{TSV,bump}$), and larger yields lowering CFP compared to smaller TSV pitches.

C. Chiplet reusability for reduced CFP

Besides the ability to "mix and match" technology nodes for different chiplets and the improvements in yield with a HI system, the ability to reuse chiplets also helps lower C_{emb} as the C_{des} and the NRE component of the C_{mfg} is amortized across the number of chiplets manufactured (N_{M_i}) and used in a variety of systems in different applications. Several standard IP blocks such as USB, PCIe etc., can be manufactured in large volumes as chiplets that can then be used across several different systems amortizing C_{des}. Further, when chiplets are manufactured in large volumes, the CFP associated with NRE costs, such as manufacturing and designing the masks used during photolithography, also gets amortized across N_{M_i}. Although ECO-CHIP does not split the C_{mfg} into its NRE and non-NRE components, this will only improve CFP savings.

Fig. 12(a) shows the sweep of the ratio of N_{M_i} to N_S for the EMR 2-chiplet testcase and plots C_{des} (with $N_{\text{iter}} = 100$,

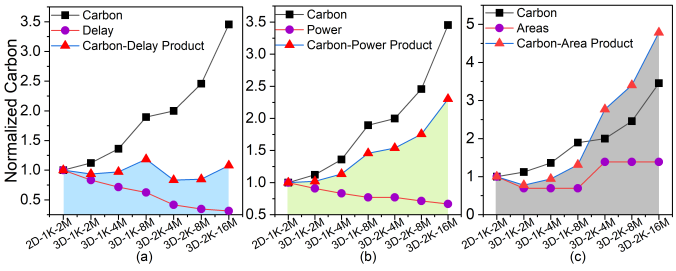

Fig. 13: (a) Carbon-delay product curve, (b) carbon-power product curve, and (c) carbon-area product for different 3D configurations of the accelerator.

$N_{M_i} = 100,000$)(Refer Eq. (12)) when both chiplets are implemented in 7nm technology node. Larger $\frac{N_{M_i}}{N_S}$ results in lower C_{des} as the cost of designing the chiplet is amortized across a larger number of systems. The figure shows the potential C_{des} savings from the reusability of chiplets, i.e., chiplets can be designed once and reused several times.

Fig. 12(b), (c), and (d) shows C_{tot} across a lifetime of 5 years for different $\frac{N_{M_i}}{N_S}$ ratios for the GA102 and A15 with RDL fanout-based packaging, and EMR with EMIB-based packaging testcases, respectively. With the increase in the ratio, the C_{emb} reduces, and with an increase in lifetime, the C_{op} increases. On one hand, in GA102 (Fig. 12(b)), the C_{op} dominates C_{emb} and therefore, C_{tot} does not reduce significantly with the increase in the ratio. On the other hand, in the A15 testcase (Fig. 12(c)), where the C_{emb} dominates the C_{op}, increasing the ratio helps lower C_{tot}. Therefore, in A15, reducing C_{tot} requires reducing the C_{emb}, especially given the lifetime of these consumer mobile processors is small. With a N_{M_i} value of 100,000, used across all testcases, the figure indicates how many systems each chiplet must be utilized in, to amortize the embodied cost across the lifetime of operation. This analysis is helpful when determining the required volumes in which chiplets much be manufactured.

VI. Design and Architecture Space Exploration for Chiplet Disaggregation

In this section, we demonstrate the application of ECO-CHIP in performing design space exploration, considering CFP as a first-order optimization metric along with performance, power, area, and cost. To analyze the CFP tradeoffs with these metrics, we consider the accelerator testcase described for AR/VR applications [55] in addition to three testcases in Section V. The testcase uses a 3D packaging integration technique with 1-4 SRAM dies stacked on top of a computation unit using microbumps in a 7nm technology. The testcase comes in two flavors. The first, 1K, uses SRAM dies of 2MB capacity each, and the second, 2K, uses SRAM dies of 4MB capacity each. The naming convention for each of these testcases is as follows: 2D/3D-1K/2K-2MB/4MB/8MB/16MB. For instance, a 3D-1K-4M is a 3D architecture with 2 tiers of a 2MB SRAM chiplet and a total memory of 4MB.

(1) Delay, power, area, and CFP tradeoffs Fig. 13(a) shows the carbon-delay product curve for the accelerator testcase with different numbers of SRAM tiers. As SRAM tiers increase (from left to right in the figure) for each 1K or 2K series, the system latency reduces, but ECO-CHIP finds that C_{tot} increases. C_{tot} is estimated for a lifetime of 2 years, where C_{op} is estimated using E_{use} provided in [55]. ECO-CHIP shows that as the number of tiers increases, although the delay improves, the embodied C_{emb} increases as there is an increase in total memory capacity and silicon dies.

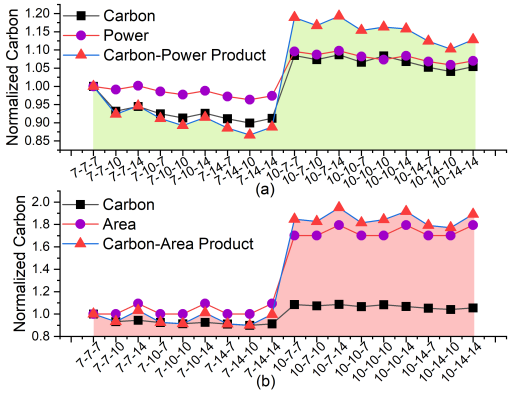

Fig. 14: (a) Carbon-power product and (b) carbon-area product for GA102.

Fig. 13(b) and (c) show the carbon-power and the carbon-area product curves for the same accelerator testcase. The curves show that as the number of SRAM tiers increases, the energy efficiency of the accelerator improves, reducing operational power [55] and carbon. However, since the C_{emb} dominates, the C_{tot} increases as the number of tiers increases. Since it's a 3D system, the 2-dimensional area of each configuration remains within 1K or 2K. These product curves enable design space exploration, allowing the selection of an architecture that meets the latency, power, and area specifications while minimizing C_{tot}.

Since the performance of the HI system is very application- and testcase-specific, estimating the performance overheads of the chiplet-based GA102 testcase and A15 testcase, which are originally monolithic, requires modeling the performance of inter-die communication and router overheads, which is beyond the scope of ECO-CHIP. Therefore, for performance and CFP tradeoff curve, we only consider the accelerator testcase with delay numbers that are readily available [55].

Fig. 14(a) and (b) show the variation in operational power and area and total CFP product for the GA102 3-chiplet with RDL fanout package testcase for different technology nodes normalized to its monolithic counterpart. Older technology node chiplets have larger chip areas and power due to HI-related overheads such as white space on the substrate/interposer, additional router logic etc. However, older technology nodes have a lower CFPA lowering C_{tot} (Fig. 7). ECO-CHIP enables considering these tradeoffs, to drive decisions related to SoC disaggregation into chiplets.

(2 Dollar cost analysis ECO-CHIP integrates with a third-party chiplet-based dollar cost analysis tool [27] and uses default parameters in [59] for cost estimation. We input the architectural description (areas and technology nodes) of our testcase with identical yield numbers used for CFP estimation. Fig. 15(a) shows the dollar cost associated with the 3-chiplet GA102 testcase for different technology node combinations.

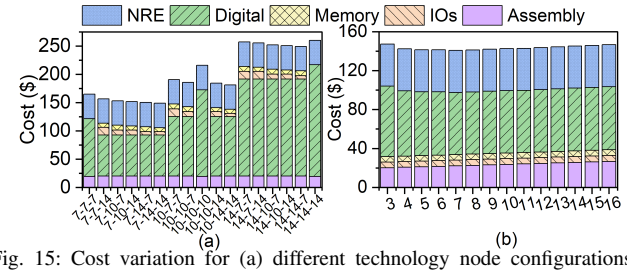

Fig. 15: Cost variation for (a) different technology node configurations of GA102 testcase, (b) disaggregating the GA102 testcase into N_C chiplets.

The dollar cost follows a similar trend as the total CFP trend in Fig. 7(d), where older technology node chiplets have lower costs due to better yields and cheaper manufacturing.

Fig. 15(b) shows the variation in dollar cost as we split the GA102 digital logic block into N_C chiplets similar to manufacturing CFP cost in Fig. 10. The assembly cost increases as the number of chiplets increases and the cost of manufacturing the digital logic block decreases due to an increase in yield of smaller die sizes as the number of chiplets increases. The cost variation in Fig. 15(b) is small compared to the variation in C_{HI} and C_{mfg} in Fig. 10, which allows an architect to consider the N_C with the least C_{tot} from a cost perspective.

VII. Validation Discussion

It's important to note here that ECO-CHIP is a tool to perform analysis on embodied and operational CFP of heterogenous (HI) systems which has not been done before. It's a methodology that is available open-source and can generate numbers as accurate as the accuracy of the input parameters, e.g., design time, yields, and defect densities, and is easily adoptable by the industry that has access to accurate numbers. ECO-CHIP is based on CFP data from two sources. First, [5] which provides manufacturing CFP numbers from IMEC reported on a per metal layers basis for different technology nodes. Second, ACT [4], which provides manufacturing CFP numbers mined from industrial sustainability reports [6], [8], [9], [33], [36]–[38], and a carbon footprint per unit area. The HI-related CFP estimation also relies on the same CFP numbers and is estimated by modeling additional area overheads and metal layers in interposer/package substrate for each kind of packaging architecture. As a sanity check of the CFP numbers generated by ECO-CHIP, we compare our A15 processor CFP numbers with that reported in Apple's sustainability report for the entire iPhone 14. We find that our reported numbers are approximately 16% of the total CFP of the iPhone. As reported by the Apple sustainability and product report [60], [61], ECO-CHIP also estimates 20% of the total CFP is for operational CFP and 80% for embodied CFP as seen in Fig. 8(b). Validation is indeed a challenging problem, especially given the coarse granularity at which industry sustainability reports are provided and the lack of open-source data from the industry on various input parameters such as design time, yields, etc. For example, Apple provides CFP for the iPhone as a whole and it's difficult to figure out the contributions of the A15 processor alone. However, the industry that has accurate data on yields, design time, etc., can utilize ECO-CHIP to generate accurate CFP.

VIII. Related Work

Two prior bodies of work focus on CFP estimation at the architectural level: the first body of work includes [19], [62], and the second includes [4], [7], [63]. The work in [19] reformulated the Kaya identity [64] to understand how the global CFP of computer systems evolves and has made a case to lower chip sizes to lower embodied CFP and [62] creates a simple model based on first principles. The works in [4], [7], [63] have created data-driven model, from publicly available sustainability reports from industry [3], [6], [8], [33], for embodied carbon estimation and have created a platform for carbon-aware design space exploration (DSE) [7]. While these works have set a new paradigm, they are limited in scope:

1) They do not apply to emerging HI systems where small chiplets are integrated into a single package.
2) They do not accurately consider the packaging/assembly carbon costs, which is crucial for HI systems. ACT [4] uses a fixed package CFP value irrespective of the size of the package, the yield of the package, and the assembly process and [19] does not consider or separate the packaging CFP component.
3) They do not consider the CFP from the *design* of chips, which, even though amortized across all manufacturing parts, significantly contributes to the embodied CFP.
4) They do not take into account silicon wastage from the periphery of the wafer (Fig. 3).

In contrast, and complementary, to the above two bodies of work ECO-CHIP focuses on evaluating the potential of chiplet-based systems towards sustainable computing by modeling CFP from advanced packaging architectures, yields, area scaling models, and design CFP. Our comparisons to [4], [7] in Section V have shown the importance of considering these models in the context of HI. Unlike [4], [7], ECO-CHIP is integrated with emerging design methodologies (such as system disaggregation) for chiplet-based systems to make them cognizant of sustainability (Refer Section VI).

IX. Conclusion

In this paper, we proposed HI as a path towards sustainable computing by designing and manufacturing chiplet-based systems with lower embodied carbon footprint (CFP) than monolithic SoCs. We developed ECO-CHIP, a CFP estimator that uses architectural-level descriptions to assess heterogeneous systems' total CFP (embodied and operational), including advanced packaging CFP overheads. We demonstrated the use of ECO-CHIP to guide system disaggregation and design space exploration in Section VI and integrated with other chiplet-based cost estimation tools. ECO-CHIP is open-source and available at anonymous repository [21]. We believe that ECO-CHIP will enable the development of more sustainable design methodologies for emerging heterogeneous systems.

REFERENCES

[1] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, and A. Friday, "The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations," *Patterns*, vol. 2, no. 9, p. 100340, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389921001884

[2] B. Graver, K. Zhang, and D. Rutherford, "CO2 EMISSIONS FROM COMMERCIAL AVIATION," 2019, https://theicct.org/publication/co2-emissions-from-commercial-aviation-2018/ (last accessed: March 2023).

[3] L. Å. Ragnarsson, C. Rolin, S. Shamuilia, and E. Parton, "The green transition of the IC industry," 2022, https://www.imec-int.com/en/expertise/cmos-advanced/sustainable-semiconductor-technologies-and-systems-ssts/stss-white-paper (last accessed: March 2023).

[4] 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 *Proceedings of the ACM International Symposium on Computer Architecture*, 2022, p. 784–799.

[5] M. Garcia Bardon, P. Wuytens, L.-r. Ragnarsson, G. Mirabelli, D. Jang, G. Willems, A. Mallik, A. Spessot, J. Ryckaert, and B. Parvais, "DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies," in *Proceedings of the IEEE International Electron Devices Meeting*, 2020, pp. 41.4.1–41.4.4.

[6] Apple, "Environmental Responsibility Report," 2019, https://www.apple.com/environment/pdf/Apple_Environmental_Responsibility_Report_2019.pdf.

[7] M. Elgamal, D. Carmean, E. Ansari, O. Zed, R. Peri, S. Manne, U. Gupta, G.-Y. Wei, D. Brooks, G. Hills, and C.-J. Wu, "Design Space Exploration and Optimization for Carbon-Efficient Extended Reality Systems," 2023.

[8] Facebook, "Facebook Sustainability Data 2019," 2019, https://sustainability.fb.com/wp-content/uploads/2020/05/2019-Sustainability-Data-Disclosure_Final-1.pdf.

[9] Microsoft, "Environmental Sustainability Report," 2021, https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report.

[10] S. Naffziger, K. Lepak, M. Paraschou, and M. Subramony, "AMD Chiplet Architecture for High-Performance Server and Desktop Products," in *Proceedings of the IEEE International Solid- State Circuits Conference*, 2020, pp. 44–45.

[11] M. T. Bohr and I. A. Young, "CMOS Scaling Trends and Beyond," *IEEE Micro*, vol. 37, no. 6, pp. 20–29, 2017.

[12] S. S. Iyer, "Heterogeneous Integration for Performance and Scaling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, no. 7, pp. 973–982, 2016.

[13] P. Gelsinger, "Keynote: Semiconductors run the world," hotChips, 2022.

[14] R. Mahajan, R. Sankman, N. Patel, D.-W. Kim, K. Aygun, Z. Qian, Y. Mekonnen, I. Salama, S. Sharan, D. Iyengar, and D. Mallik, "Embedded multi-die interconnect bridge (EMIB) – a high density, high bandwidth packaging interconnect," in *Proceedings of the IEEE Electronic Components and Technology Conference*, 2016, pp. 557–565.

[15] S. Y. Hou, W. C. Chen, C. Hu, C. Chiu, K. C. Ting, T. S. Lin, W. H. Wei, W. C. Chiou, V. J. C. Lin, V. C. Y. Chang, C. T. Wang, C. H. Wu, and D. Yu, "Wafer-Level Integration of an Advanced Logic-Memory System Through the Second-Generation CoWoS Technology," *IEEE Transactions on Electron Devices*, vol. 64, no. 10, pp. 4071–4077, 2017.

[16] D. B. Ingerly, S. Amin, L. Aryasomayajula, A. Balankutty, D. Borst, A. Chandra, K. Cheemalapati, C. S. Cook, R. Criss, K. Enamul, W. Gomes, D. Jones, K. C. Kolluru, A. Kandas, G.-S. Kim, H. Ma, D. Pantuso, C. Petersburg, M. Phen-givoni, A. M. Pillai, A. Sairam, P. Shekhar, P. Sinha, P. Stover, A. Telang, and Z. Zell, "Foveros: 3D Integration and the use of Face-to-Face Chip Stacking for Logic Devices," in *Proceedings of the IEEE International Electronic Devices Meeting*, 2019, pp. 19.6.1–19.6.4.

[17] Y. H. Chen, C. A. Yang, C. C. Kuo, M. F. Chen, C. H. Tung, W. C. Chiou, and D. Yu, "Ultra High Density SoIC with Sub-micron Bond Pitch," in *Proceedings of the IEEE Electronic Components and Technology Conference*, 2020, pp. 576–581.

[18] IEEE Components and Packaging Society, ""Heterogeneous Integration Roadmap"," 2021, http://eps.ieee.org/hir.

[19] L. Eeckhout, "Kaya for Computer Architects: Toward Sustainable Computer Systems," *IEEE Micro*, vol. 43, no. 1, pp. 9–18, 2023.

[20] Y. Feng and K. Ma, "Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration," in *Proceedings of the ACM/IEEE Design Automation Conference*, ser. DAC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 121–126. [Online]. Available: https://doi.org/10.1145/3489517.3530428

[21] "ECO-CHIP," 2023, https://github.com/ASU-VDA-Lab/ECO-CHIP.

[22] W. Gomes, S. Morgan, B. Phelps, T. Wilson, and E. Hallnor, "Meteor Lake and Arrow Lake Intel Next-Gen 3D Client Architecture Platform with Foveros," in *Proceedings of the IEEE Hot Chips Symposium*, Aug. 2022, pp. 1–40. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/HCS55958.2022.9895532

[23] J. Ferguson, "EDA innovation is the foundation of progress," https://www.techdesignforums.com/practice/technique/physical-verification-eda-innovation-is-the-foundation-of-progress/.

[24] P. Ehrett, T. Austin, and V. Bertacco, "Chopin: Composing Cost-Effective Custom Chips with Algorithmic Chiplets," in *Proceedings of the IEEE International Conference on Computer Design*, 2021, pp. 395–399.

[25] J. H. Lau, "Recent Advances and Trends in Advanced Packaging," *IEEE Transactions on Components and Packaging Technologies*, vol. 12, no. 2, pp. 228–252, 2022.

[26] A. B. Kahng, B. Lin, and S. Nath, "ORION3.0: A Comprehensive NoC Router Estimation Tool," *IEEE Embedded Systems Letters*, vol. 7, no. 2, pp. 41–45, 2015.

[27] A. Graening, S. Pal, and P. Gupta, "Chiplets: How Small is too Small?" in *Proceedings of the ACM/IEEE Design Automation Conference*, July 2023.

[28] Angstronomics, "The Truth of TSMC 5nm," https://www.angstronomics.com/p/the-truth-of-tsmc-5nm.

[29] D. Schor, "IEDM 2022: Did We Just Witness The Death Of SRAM?" https://fuse.wikichip.org/news/7343/iedm-2022-did-we-just-witness-the-death-of-sram/.

[30] J. 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.

[31] A. Ning, G. Tziantzioulis, and D. Wentzlaff, "Supply Chain Aware Computer Architecture," in *Proceedings of the ACM International Symposium on Computer Architecture*, 2023.

[32] D. Stow, Y. Xie, T. Siddiqua, and G. H. Loh, "Cost-effective design of scalable high-performance systems using active and passive interposers," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, 2017, pp. 728–735.

[33] TSMC, "Corporate Social Responsibility Report," 2018, https://esg.tsmc.com/download/file/2018_tsmc_csr_report_published_May_2019/english/pdf/e_all.pdf.

[34] T. Li, J. Hou, J. Yan, R. Liu, H. Yang, and Z. Sun, "Chiplet Heterogeneous Integration Technology—Status and Challenges," *Electronics*, vol. 9, no. 4, 2020.

[35] C.-H. Kuo, A. Hu, L. Hung, K.-T. Yang, and C.-H. Wu, "Life cycle impact assessment of semiconductor packaging technologies with emphasis on ball grid array," *Journal of Cleaner Production*, vol. 276, p. 124301, 12 2020.

[36] SPIL, "Corporate Social Responsibility Report," 2018, https://www.spil.com.tw/Files/pdf-en/2018-en.pdf.

[37] Amkor Technology Inc., "SUSTAINABILITY ACCOUNTING STANDARDS BOARD REPORT," 2021, https://amkor.com/esg/esg-report/.

[38] ACE Group, "Corporate sustainability report," 2017, https://ase.aseglobal.com/en/csr/downloads/report.

[39] R. Mahajan, Z. Qian, R. S. Viswanath, S. Srinivasan, K. Aygün, W.-L. Jen, S. Sharan, and A. Dhall, "Embedded Multidie Interconnect Bridge—A Localized, High-Density Multichip Packaging Interconnect," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 10, pp. 1952–1962, 2019.

[40] Z. Cheng, Y. Ding, Z. Zhang, M. Zhou, and Z. Chen, "Coupled Thermo-Mechanical Analysis of 3D ICs Based on an Equivalent Modeling Methodology With Sub-Modeling," *IEEE Access*, vol. 8, pp. 14 146–14 154, 2020.

[41] S. W. Liang, G. C. Y. Wu, K. C. Yee, C. T. Wang, J. J. Cui, and D. C. H. Yu, "High performance and energy efficient computing with advanced soic™ scaling," in *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, 2022, pp. 1090–1094.

[42] D. Stow, I. Akgun, and Y. Xie, "Investigation of Cost-Optimal Network-on-Chip for Passive and Active Interposer Systems," in *Proceedings of*

the ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP), 2019, pp. 1–8.
[43] C. Sun, C.-H. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic, "DSENT - A Tool Connecting Emerging Photonics with Electronics for Opto-Electronic Networks-on-Chip Modeling," in *International Symposium on Networks-on-Chip*, 2012, pp. 201–210.
[44] C. J. Alpert, D. P. Mehta, and S. S. Sapatnekar, *Handbook of Algorithms for Physical Design Automation*, 1st ed. USA: Auerbach Publications, 2008.
[45] S. Y. Hou, W. C. Chen, C. Hu, C. Chiu, K. C. Ting, T. S. Lin, W. H. Wei, W. C. Chiou, V. J. C. Lin, V. C. Y. Chang, C. T. Wang, C. H. Wu, and D. Yu, "Wafer-Level Integration of an Advanced Logic-Memory System Through the Second-Generation CoWoS Technology," *IEEE Transactions on Electron Devices*, vol. 64, no. 10, pp. 4071–4077, 2017.
[46] I. Cutress, "'Better Yield on 5nm than 7nm': TSMC Update on Defect Rates for N5," https://www.anandtech.com/show/16028/better-yield-on-5nm-than-7nm-tsmc-update-on-defect-rates-for-n5.
[47] J. 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.
[48] Y. Chen, D. Niu, Y. Xie, and K. Chakrabarty, "Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis," in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2010, pp. 471–476.
[49] F450C, "Evolution of the Silicon Wafer," https://f450c.org/infographic/.
[50] "List of CPU power dissipation figures," https://en.wikipedia.org/wiki/List_of_CPU_power_dissipation_figures.
[51] P. Agrawal, M. Broxterman, B. Chatterjee, P. Cuevas, K. H. Hayashi, A. B. Kahng, P. K. Myana, and S. Nath, "Optimal Scheduling and Allocation for IC Design Management and Cost Reduction," *ACM Transactions on Design Automation of Electronic Systems*, vol. 22, no. 4, jun 2017. [Online]. Available: https://doi.org/10.1145/3035483
[52] D. Patel, G. Wong, G. Cozma, and Locuza, "Intel Emerald Rapids Backtracks on Chiplets – Design, Performance & Cost," https://www.semianalysis.com/p/intel-emerald-rapids-backtracks-on.
[53] Z. Liu, "IR Photographer Shares Die Shots of Nvidia 3000 Series GA102 Silicon," https://www.tomshardware.com/news/infrared-photographer-photos-nvidia-ga102-ampere-silicon.
[54] Wikipedia, "Apple A15 Processor," https://en.wikipedia.org/wiki/Apple_A15.
[55] L. Yang, R. M. Radway, Y.-H. Chen, T. F. Wu, H. Liu, E. Ansari, V. Chandra, S. Mitra, and E. Beigné, "Three-Dimensional Stacked Neural Network Accelerator Architectures for AR/VR Applications," *IEEE Micro*, vol. 42, no. 6, pp. 116–124, 2022.
[56] D. Patel, "Apple A15 Die Shot and Annotation – IP Block Area Analysis," https://www.semianalysis.com/p/apple-a15-die-shot-and-annotation.
[57] P. Alcorn, "Intel Details Tiger Lake at Hot Chips 2020, Die Revealed," https://www.tomshardware.com/news/intel-details-tiger-lake-at-hot-chips-2020-die-revealed.
[58] NVIDIA, "NVIDIA Ampere GA102 GPU Architecture," 2021, https://images.nvidia.com/aem-dam/en-zz/Solutions/geforce/ampere/pdf/NVIDIA-ampere-GA102-GPU-Architecture-Whitepaper-V1.pdf.
[59] A. Graening, S. Pal, and P. Gupta, "Cost model chiplets," https://github.com/nanocad-lab/cost_model_chiplets.
[60] Apple, "Environmental Responsibility Report," 2023, https://www.apple.com/environment/pdf/Apple_Environmental_Progress_Report_2023.pdf.
[61] ——, "Product Environmental Report," 2022, https://www.apple.com/environment/pdf/products/iphone/iPhone_14_and_iPhone_14_Plus_PER_Sept2022.pdf.
[62] L. Eeckhout, "A First-Order Model to Assess Computer Architecture Sustainability," *IEEE Computer Architecture Letters*, vol. 21, no. 2, pp. 137–140, 2022.
[63] U. Gupta, Y. 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," *IEEE Micro*, vol. 42, no. 4, p. 37–47, jul 2022. [Online]. Available: https://doi.org/10.1109/MM.2022.3163226
[64] "Kaya Identity," https://en.wikipedia.org/wiki/Kaya_identity.
[65] C. Choppali Sudarshan, N. Matkar, S. Vrudhula, S. Sapatnekar, and V. Chhabria, "ECO-CHIP: Estimation of Carbon Footprint of Chiplet-based Architectures for Sustainable VLSI: HPCA 2024 Artifact Evaluation," Nov. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10223759

APPENDIX

A. Abstract

The paper introduces ECO-CHIP, a framework for measuring the carbon footprint (CFP) of a heterogeneous system across its lifespan. This artifact is released on Zenodo and contains two parts. The first is ECO-CHIP submodule from GitHub, and the second is a folder that consists of the experiments performed using ECO-CHIP. This appendix describes the installation of our artifact, ECO-CHIP, and the procedure to reproduce the results in the paper. The minimal hardware requirements are any single-core CPU, and the software requirements are Python 3.8, python3.8-venv, with pip 20.0.2.

B. Artifact check-list (meta-information)

- **Algorithm:** Heterogeneous chiplet based carbon analysis tool that can evaluate the sustainability potential of Heterogeneous systems, considering scaling, chiplet, packaging yields, design complexity, and advanced packaging overheads.
- **Program:** ECO-CHIP and the experiments are setup in Python.
- **Compilation:** Python compiler
- **Data set:** There is no large particular dataset. However, in our ECO-CHIP GitHub repository [21] and the Zenodo release [65], we have configuration files and architectural descriptions of the testcases used in the paper, which serve as input to ECO-CHIP in JSON format.
- **Run-time environment:** Runtime is not critical for ECO-CHIP. The script performs very simple equation-based calculations.
  - Not OS-Specific
  - Dependencies - Python 3.8, python3.8-venv, pip 20.0.2
  - No need for root access
- **Hardware:** No specific hardware requirement, at least one CPU core.
- **Run-time state:** Not sensitive to run-time state.
- **Execution:** Full execution should take 10sec based on input parameters.
- **Metrics:** ECO-CHIP simulator estimates equivalent CO_2 emissions. The output will be the total CFP for the input testcase, including embodied CFP (design and manufacturing) and operational CFP.
- **Output:**
  - ECO-CHIP will output the CO_2 emission values across different combinations of technology nodes.
  - Will provide a breakdown of CO_2 values across different chiplets.
  - Provides design and manufacture (embodied), operational, and total CO_2 emission values.
- **Experiments:** The experimental setup that we release on Zenodo (artifact) generates the key results of the paper (Fig. 7, 8, 9, 10, 13, etc.). The experimental setup is specific to the results in the paper, where the testcase sweeps of various input parameters, etc., are present in the specific directory related to the figure to help with easy reproducibility. ECO-CHIP GitHub is set up to run different new testcases (not just the ones in the paper) and estimate CFP of the system. We provide the experiment scripts and detailed descriptions for running them. All instructions are provided in `README.md` file under `artifact` available on Zenodo (Please note that this is not available on GitHub).
- **How much disk space required?:** Less than 1GB
- **How much time is needed to prepare workflow (approximately)?:** Less than 1 minute
- **How much time is needed to complete experiments (approximately)?:** Less than 10 minutes

- **Publicly available?:** Yes.
  - https://github.com/ASU-VDA-Lab/ECO-CHIP
  - https://zenodo.org/records/10223759
- **Code licenses (if publicly available)?:** BSD 3-Clause "New" or "Revised" License
- **Archived (provide DOI)?:**
  - https://github.com/ASU-VDA-Lab/ECO-CHIP
  - Zenodo DOI : 10.5281/zenodo.10099731

C. Description

1) How to access: The artifact to regenerate all the results in the paper is available on open-source Zenodo [65]. ECO-CHIP simulator is available on GitHub [21].

2) Hardware dependencies: A CPU with at least one core.

3) Software dependencies: The artifact and the tool are implemented in Python 3.8 and require several packages that help run the tool. A full detailed list of required packages is in `requirments.txt` file. The `requirments.txt` is available in the repository.

4) Data sets: Our GitHub repository [21] contains testcases that were used in paper, most of the files are in JSON format. Detailed descriptions of each of these files are provided in `README.md` file under the repository. The `architecture.json` contains high-level architecture details of each chiplet and the packaging type, `designC.json` contains input parameters needed for design CFP, `node_list.txt` specifies the technology nodes of interest for CFP exploration, `operationalC.json` specifies details about the lifetime and `packageC.json` has specific parameters related to packaging. The description of each testcase in our dataset is provided in [21].

D. Installation

The installation for ECO-CHIP simulator and all the experiments in the artifact repository are the same. Once the zip file is downloaded from Zenodo [65] or cloned from GitHub [21], it requires creating a virtual Python environment to install all the packages via pip. Based on the following instructions:

- `cd ECO-CHIP-AE` or `cd ECO-CHIP`
- `python3 -m venv eco-chip`
- `source eco-chip/bin/activate`
- `pip3 install -r requirements.txt`

The `source eco-chip/bin/activate` assumes using bash shell on the Unix environment.

E. Experiment workflow

Experiments to regenerate results in the paper The artifact directory available on Zenodo contains scripts and details on regenerating the results in the paper. The README file in each folder details how to run each experiment. To run all experiments, after installation (as described in the top-level README), from the artifact folder, run the bash script `run_all.sh` in the virtual environment. A specific experiment can be run from its unique folder. For example,

- `cd artifact/fig2`
- `python3 fig2a.py`

The result with the plot generated will be created in the `artifact/result_img/` directory.

ECO-CHIP simulator stand-alone With the simulator [21], we provide multiple examples of testcases to measure the CFP of different heterogeneous systems and perform CFP estimation across different technology node combinations. For example, after installation, from the ECO-CHIP folder, the following command will run the GA102 testcase:

- `python3 src/ECO_chip.py --design_dir testcases/GA102/.`

We have multiple other testcases under the `ECO-CHIP/testcases` directory.

Running ECO-CHIP for a new design To run ECO-CHIP on a new design, create a new directory under `ECO-CHIP/testcase/` and add the parameters of the design of interest into the JSON files. Detailed description for each of the input parameter files is provided in `ECO-CHIP/README.md` and then run the above command with a pointer to the specific design directory.

F. Evaluation and expected results

All the required scripts to help reproduce the key results and contributions are under the `artifact` directory. Detailed steps on how to run each of the scripts and Python packages that need to be installed have been mentioned in the `README.md` file. The results of our artifact will be reproducible and may have very small variations in single-digit percentages. All the graphs generated by the script will be under `artifact/result_img/` directory. Running the scripts such as `fig7.py`, `fig8a.py`, `fig8b.py`, `fig9.py`, `fig10.py`, and `fig13.py` will help generate the respective plots under the `artifact/result_img/` directory. This can verify the critical result of the paper as shown in Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 13, as the main contribution in using chiplet-based systems as a sustainable alternatives to monolithic designs. In addition to generating the plot, the script also prints the underlying raw data within the plot.

G. Experiment customization

The experiment can be customized based on the input parameters that are provided to ECO-CHIP, `architecture.json` can customize the system architecture stating the chiplet sizes and types, along with package type that is used, in-depth packaging parameters can be customized under `packageC.json`. Parameters under `design.json` can be customized to explore more on the design CFP. Customizing `node_list.txt` can help in exploring across different nodes. All the details for each of the input parameter files are explained under `ECO-CHIP/README.md`.

H. Methodology

Submission, reviewing, and badging methodology:

- https://www.acm.org/publications/policies/artifact-review-badging
- http://cTuning.org/ae/submission-20201122.html
- http://cTuning.org/ae/reviewing-20201122.html