2025 Design, Automation & Test in Europe Conference (DATE 2025) – Best Paper Award Candidate

Quantifying Trade-Offs in Power, Performance, Area, and Total Carbon Footprint of Future Three-Dimensional Integrated Computing Systems

Danielle Grey-Stewart†*, David Kong†*, Mariam Elgamal, Georgios Kyriazidis, Jalil Morris, and Gage Hills
Harvard School of Engineering and Applied Sciences (SEAS)
†Emails: dgreystewart@fas.harvard.edu, dkong@g.harvard.edu *Equal contributions

***Abstract*—To address computing's carbon footprint challenge, designers of computing systems are beginning to consider *carbon footprint* as a first-class figure of merit, alongside conventional metrics such as power, performance, and area. To account for *total carbon* (tC) *footprint* of a computing system, carbon footprint models must consider both *embodied carbon* ($C_{embodied}$) due to emissions during manufacturing, and *operational carbon* ($C_{operational}$) from day-to-day use. Models for $C_{operational}$ are relatively mature due to the direct relationship between $C_{operational}$ and energy consumed while computing. In contrast, models for $C_{embodied}$ primarily focus on today's *silicon*-based technologies, not capturing the wide range of *beyond-Si* technologies that are actively being developed for future computing systems, including emerging nanomaterials, emerging memory devices, and various three-dimensional (3D) integration techniques. $C_{embodied}$ models for emerging technologies are essential for accurately predicting which technology directions to pursue without exacerbating computing's carbon footprint.**

In this paper, we (1) develop $C_{embodied}$ models for 3D-integrated computing systems that leverage emerging nanotechnologies. We analyze an example fabrication process that is highly promising for energy-efficient computing: 3D integration of carbon nanotube field-effect transistors (CNFETs) and indium gallium zinc oxide (IGZO) FETs fabricated directly on top of Si CMOS at a 7 nm technology node. We show that $C_{embodied}$ of this process is, on average (considering various energy grids), 1.31× higher per wafer vs. a baseline 7 nm node Si CMOS process. (2) As a case study, we quantify trade-offs in power, performance, area, and tC footprint for an embedded system comprising an ARM Cortex-M0 processor and embedded DRAM, implemented in each of the above processes. For a representative lifetime of the system (running applications from the Embench suite for 2 hours per day over 24 months, with a clock frequency of 500 MHz), we show that the 3D IGZO/CNFET/Si implementation is 1.02× more *carbon-efficient* per good die (considering yield) vs. the baseline Si implementation, quantified by the product of tC and application execution time (*tCDP*, an effective metric of carbon efficiency). (3) Finally, we show techniques to quantify carbon efficiency benefits of future computing systems, even when there is uncertainty in carbon footprint models. Specifically, we show how to robustly compare tCDP for multiple computing systems, given underlying uncertainty in $C_{embodied}$, computing system lifetime, carbon intensity (in equivalent grams of CO_2 emissions per unit energy consumption), and yield.

I. Introduction

The carbon footprint of the Information and Communication Technology (ICT) sector is estimated to comprise up to 4% of global carbon emissions, quantified in mass equivalents of carbon dioxide (CO_2e) [1]–[3]. This footprint is increasing rapidly with the explosive growth in training large language models and in creating embedded systems for consumer electronics. To address this global challenge, designers are beginning to consider carbon footprint as a key figure of merit for computing systems, alongside conventional metrics such as power, performance, and area [4]–[6].

Importantly, *total* carbon (tC) footprint accounts for the total emissions of a computing system across its entire lifetime, including both *operational carbon* ($C_{operational}$) and *embodied carbon* ($C_{embodied}$) [6]. $C_{operational}$ is quantified using industry-standard Electronic Design Automation (EDA) tools for modeling power/energy consumption, combined with public data that reports carbon intensity during operation (CI_{use}, in units of grams CO_2e per unit energy consumption). Conversely, $C_{embodied}$ models are relatively immature as they must account for emissions during fabrication (due to energy consumed by equipment, material and gas usage, and carbon intensity during fabrication, CI_{fab}), which are challenging to quantify [5].

Fortunately, researchers are actively developing $C_{embodied}$ models for state-of-the-art technology nodes. While these models are valuable, they primarily rely on top-down life cycle analyses [7]–[9] or focus on solely on *silicon*-based technologies [4], and thus do not capture the wide range of *beyond-Si* technologies currently being investigated for future directions in energy-efficient computing. New directions in computing technologies include emerging nanomaterials (carbon nanotubes [10], two-dimensional materials [11]), memory devices (resistive RAM [12], embedded DRAM (eDRAM) using IGZO [13]), and various 3D integration techniques such as integrating chiplets onto interposers [7], [14], and *monolithic 3D integration (M3D)*, in which multiple layers of computing and memory devices are fabricated sequentially over the same starting substrate [15], [16]. It is critical to understand the implications of these technologies on computing's carbon footprint as they are being developed.

As an example, consider an M3D integrated circuit (IC) that has multiple layers of computing and memory circuits, which are densely connected using fine-grained vertical interconnects. The circuits in the bottom layer are implemented using Si CMOS, and the upper circuit layers are implemented using emerging nanomaterials (we describe and analyze such M3D ICs extensively in Sections **II** and **III**). M3D ICs offer a key trade-off for carbon-efficient computing systems:

- *Energy efficiency benefit:* the *memory wall* is a major system-level bottleneck for today's computing systems, where the vast majority of application execution time and energy consumption is spent passing data between the processors and off-chip memory [15]. M3D ICs permit massive amounts of on-chip memory to be integrated directly on top of processors, enabling low latency memory accesses and high memory bandwidth. The resulting energy efficiency benefits translate directly into $C_{operational}$ benefits.
- *$C_{embodied}$ drawback:* M3D IC fabrication requires more process steps to realize additional layers of transistors, memory, and metal routing layers (Sec. **II**-C), leading to higher $C_{embodied}$ per wafer.

Thus, it is unclear whether M3D ICs should be pursued for carbon-efficient computing when considering trade-offs in power, performance, area, and total carbon (PPAtC). A robust

978-3-9826741-0-0/DATE25/© 2025 EDAA

framework that accurately quantifies tC of future computing systems is essential for identifying which technologies should be pursued to overcome computing's carbon footprint challenge. Toward this goal, our key contributions are:

1) We develop $\mathrm{C_{embodied}}$ models for 3D-integrated computing systems that leverage emerging nanotechnologies (Sec. **II**).
2) Using these models, we analyze an example process technology that is highly promising for energy-efficient computing: monolithic 3D integration of high-performance carbon nanotube field-effect transistors (CNFETs) and low-power indium gallium zinc oxide (IGZO) FETs fabricated directly on top of Si CMOS at a 7 nm node. The embodied footprint of this process is, on average (considering various energy grids), 1.31× higher per 300 mm wafer vs. a baseline 7 nm node Si CMOS process (Sec. **II**).
3) We quantify trade-offs in PPAtC for an embedded system comprising an ARM Cortex-M0 processor and eDRAM, implemented in each of the processes shown in Figure 1 (details in Sec. **III**). Over a representative lifetime (running applications from the Embench benchmark suite [17] for 2 hours per day over 24 months, with a core clock frequency of 500 MHz), we show that the 3D IGZO/CNFET/Si implementation is 1.02× more *carbon-efficient* (quantified by the product of tC and application execution time: *tCDP* [18]) per die vs. the baseline Si implementation. We show how to quantify tCDP of each system, by combining our tC models with detailed circuit simulations of physical layouts created using standard EDA tools (Sec. **III**-C). We also show how to robustly compare tCDP of each system, given underlying uncertainty in $\mathrm{C_{embodied}}$, system lifetime, yield, and $\mathrm{CI_{use}}$.

A step-by-step guide to reproduce all results in this paper is provided in our Github repository[1].

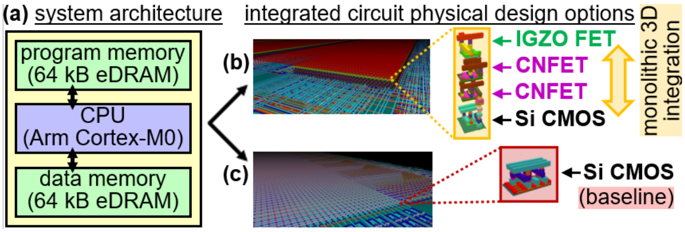
(a) system architecture
integrated circuit physical design options
program memory (64 kB eDRAM)
CPU (Arm Cortex-M0)
data memory (64 kB eDRAM)
(b)
(c)
IGZO FET
CNFET
CNFET
Si CMOS
monolithic 3D integration
Si CMOS
(baseline)

Fig. 1. Case study in comparing PPAtC of embedded computing systems. **(a)** System architecture. (b)-(c) Cross-sections of the physical layout for the M3D IGZO/CNFET/Si design (b) vs. the all-Si design (c). The ARM Cortex-M0 is implemented in Si CMOS in both cases. The eDRAM is implemented using either: M3D integration of IGZO/CNFETs/Si (b), or only Si FETs (c). Power, performance, area, and tC (PPAtC) results for each system are in Table II.

II. Modeling tC of Emerging Technologies

A. Emerging FET Technologies

M3D integration imposes processing constraints on devices that are fabricated on upper circuit layers (i.e., in the back-end-of-line or BEOL). Specifically, upper circuit layer devices must be fabricated at *low processing temperatures* (e.g., <300 °C) to avoid damaging circuits and metal interconnects fabricated in previous process steps on the wafer [15]. Fabricating today's state-of-the-art Si CMOS FETs in the BEOL would therefore be challenging as they require high temperature processing (e.g., >1,000 °C for dopant activation and annealing) [22]. Instead, a wide range of beyond-Si devices can be fabricated at low processing temperatures, enabling M3D ICs and their associated energy efficiency benefits [15].

We focus on M3D ICs leveraging two specific emerging FETs – IGZO FETs and CNFETs – integrated in the BEOL on top of Si CMOS at a 7 nm technology node. This technology enables a highly energy-efficient implementation of the embedded system in Figure 1. Notably, we design and optimize an M3D-eDRAM circuit implemented using a *combination* of Si CMOS, IGZO FETs, and CNFETs designed to leverage the benefits of each FET type while overcoming their inherent challenges (Table I). Additional details are in Section **III**-A.

TABLE I
Summary of FET benefits (+) and challenges (–).

CNFET	(+) high I_{EFF} (–) subject to metallic CNTs [10] (increases I_{OFF}) (+) BEOL-compatible (low temperature fabrication)
IGZO FET	(–) low I_{EFF} (due to low mobility) (+) ultra-low I_{OFF} (+) BEOL-compatible (low temperature fabrication)
Si FET	(+) high I_{EFF} (+) low I_{OFF} (–) bottom layer only (high temperature fabrication)

Our resulting system achieves high memory density, high data retention time, high endurance, low access latency, low access energy, and low static leakage power simultaneously. IGZO has a wide bandgap (E_g ≈3.5 eV), enabling IGZO FETs (NMOS) to exhibit ultra-low off-state leakage current (I_{OFF}) when the gate-to-source voltage (V_{GS}) is significantly below the threshold voltage (V_T). Low I_{OFF} enables IGZO eDRAM bit cells to have high data retention times (>1,000 seconds has been shown experimentally [23]), enabling low power eDRAM [23]. However, IGZO has lower carrier mobility than Si (typically <100 cm/s per V/cm [24], [25]), leading to low IGZO FET effective drive current (I_{EFF}). CNFETs exhibit high I_{EFF}, enabling high-performance circuits [26]. CNFETs typically have higher I_{OFF} than IGZO FETs, since: (1) target carbon nanotube (CNT) diameter for energy-efficient circuits is between 1 to 2 nm [26], corresponding to $0.85\text{ eV} \gtrapprox E_g \gtrapprox 0.43\text{ eV}$ [27] (low E_g limits I_{OFF}) and (2) CNTs are subject to metallic CNTs (with $E_g \approx 0$) [28], which increase I_{OFF} if they are not removed from the circuit [29].

B. Computing Total Carbon

Total carbon footprint is the sum of $\mathrm{C_{operational}}$ and $\mathrm{C_{embodied}}$. We quantify $\mathrm{C_{operational}}$ and $\mathrm{C_{embodied}}$ as follows [5], [6]. P is the power consumption during operation, and t_{life} is the system lifetime. $\mathrm{C_{embodied}}$ (to fabricate the designs shown in Figure 2a/b) is quantified in units of grams $\mathrm{CO_2}e$ ($\mathrm{gCO_2}e$).

$$\mathrm{C}_{operational} = \int_0^{t_{life}} CI_{use}(t)P(t)dt \qquad (1)$$

$$C_{embodied} = (MPA + GPA + CI_{fab} \cdot EPA) \cdot Area \qquad (2)$$

MPA is the carbon emissions per area incurred by materials procurement. We let MPA = 500 $\mathrm{gCO_2}e/\mathrm{cm}^2$, which corresponds to 3.5×10^5 $\mathrm{gCO_2}e$ per 300 mm wafer, and represents

[1] https://github.com/Harvard-NanoDesign/DATE25_PPAtC

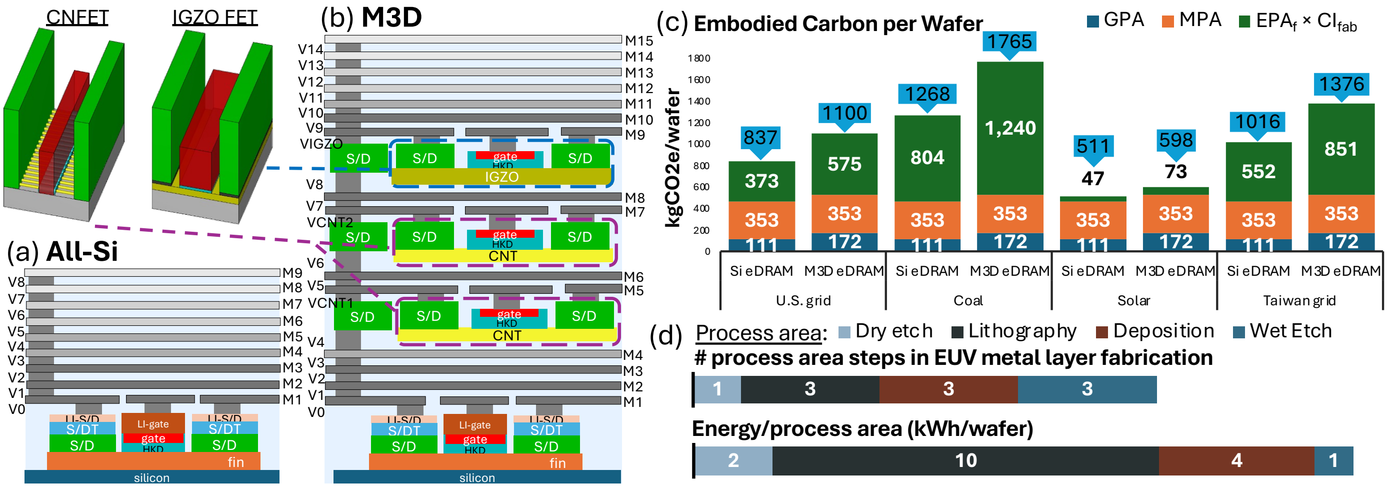

Fig. 2. **(a)** Cross section for the all-Si process, modeled after [19]. **(b)** Cross section for the M3D process with 2 tiers of CNFETs and 1 tier of IGZO FETs on top of Si CMOS. Insets show the IGZO FET and CNFET device structures. S/D(T): source/drain (trench); LI: local interconnect; HKD: high-k dielectric. **(c)** Embodied carbon *per wafer* for all-Si vs. M3D design for various power grids: U.S. (380 gCO_2e/kWh), coal (CI_{fab}=820 gCO_2e/kWh), solar (48 gCO_2e/kWh), and Taiwanese (563 gCO_2e/kWh) power grids [4], [20]. 40% electrical energy overhead is included to approximate facility energy (EPA_f = EPA×1.4), as estimated by the 2015 ITRS [21] and described at the end of Sec. **II**-C. **(d)** Steps in EUV metal layer fabrication and their total energy (kWh/wafer), adapted from reference [4]. Our Github repository includes a circuit layout (GDS) using the M3D process, with instructions on how to render it in 3D (using GDS3D).

the carbon footprint of a Si wafer as reported in life cycle analyses (LCA) [30]. Based on the carbon footprint of CNTs quantified using LCAs [31], we calculate the MPA of CNTs by multiplying the mass of CNTs deposited by the energy of CNT synthesis (≈14 $kgCO_2e$ per gram CNT [31]) when averaged across different synthesis methods). The total CNT mass per wafer in our design is on the order of picograms. Similar carbon accounting and LCA methods are needed for IGZO.

GPA refers to the carbon emissions per area incurred by direct gas emissions during fabrication. Several gases with high global warming potential (e.g., NH_3, CH_4, N_2O) are necessary inputs for fabrication processes such as etching and deposition [4]. GPA is calculated based on the mass of gas abated, data on which is available for the full processes of several technology nodes [4]. To estimate GPA for each process, both of which are modeled using a 7 nm node process design kit (PDK) [19], we use Equation 3. It scales the reported GPA from the imec iN7 EUV-patterned 7 nm node on a 300 mm wafer (0.20 $kgCO_2e/cm^2$) [4] by the ratio of fabrication energy between the iN7 node and the processes here (1.22× for the M3D process, 0.79× for the all-Si process).

$$GPA_{process} = \frac{EPA_{process}}{EPA_{iN7-EUV}} \cdot GPA_{iN7-EUV} \qquad (3)$$

C. Embodied Carbon of Fabrication Processes

EPA is the electrical energy per area incurred by fabrication. This calculation requires detailed analysis of the fabrication processes in Figure 2a/b. We begin with the front-end-of-line (FEOL) of each process, encompassing the Si FinFETs. To give an accurate approximation of the fabrication energy for the Si FinFET layers (for both 7 nm node processes in Fig. 2), we equate the FEOL fabrication energy of both processes to the front- and middle-of-line fabrication energy of the imec iN7 EUV-patterned 7 nm node reported in reference [4] (436 kWh/wafer). Next, we illustrate how we model the carbon footprint of the BEOL for the all-Si and M3D processes.

First, we describe the *process flow* for the BEOL in each case. For the all-Si process (Fig. 2a), the BEOL consists of 9 metal layers (M1-M9). Following the ASAP7 PDK [19]: M1-M3 are at 36 nm pitch, M4-M5 are at 48 nm pitch, M6-M7 are at 64 nm pitch, and M8-M9 are at 80 nm pitch. The M3D process (Fig. 2b) is the same as the all-Si process from M1 to M4. After M4, an oxide layer is deposited, upon which a CNT layer is deposited using the wet processing incubation method (≈2 nm thick [26], [32]). Next, the active region of the CNFETs is patterned and dry etched with an O_2 plasma. The source/drain electrodes are patterned and deposited (40 nm thick, the same as the source/drain layers in the ASAP7 PDK [19]). Afterwards, high-k dielectric (2 nm thick [26]) is deposited. Gate metal (30 nm gate length [19]) is then patterned and deposited. A wet etch is used to expose the source/drain. Vias are integrated on top of the gate, source, and drain layers to make connections to the 36 nm pitch metal layer on top (M5). A via (V5), metal layer at 36 nm pitch (M6), and another via (V6) are added before the next CNFET tier, with more metal/via pairs on top. The tier of IGZO FETs (IGZO 10 nm thick [33]) is fabricated using similar process steps, though IGZO deposition is achieved using Radio Frequency (RF) sputtering, and the active region is patterned with a wet etch [34]–[36]. After the two 36 nm pitch metal layers above the IGZO FETs (M9-M10), five additional metal layers are fabricated (M11-M15) at the same dimensions as M5-M9 in the all-Si process.

We now describe how to *quantify EPA* for these BEOL processes. Information on the energy consumption of fabrication processes is limited, but the fabrication energy of a metal/via pair – given its pitch and lithography method – is available [4]. We use this information to quantify the energy of fabricating metal/via pairs at 36, 64, and 80 nm pitch. For layers with 48 nm pitch, we use the fabrication energy of a metal layer with 42 nm pitch (shown in Equation 4). For example, the energy consumption of a metal/via pair at 36 nm pitch is used to model the fabrication energy of M1 and V0, M5 and VCNT1, and IGZO source/drain and V8 (Figure 2a/b, Equation 4).

To model the fabrication energy of CNFETs and IGZO FETs, we classify each required process *step* as one of the following process *areas*: dry etch, lithography, metallization, metrology, wet etch, or deposition [4]. This is convenient because reference [4] reports, for the fabrication of metal layers using a given lithography method: (1) the number of process steps used in each process area; and (2) the total fabrication energy per wafer incurred by each process area (Fig. 2d). From this information, we can estimate the fabrication energy of each process step (according to its process area) by dividing the total fabrication energy incurred by that process area by the number of times that process area is used. For example, Figure 2d shows that for metal layer fabrication using EUV lithography, there are 3 "deposition" process steps, with total fabrication energy for this process area equal to 4 kWh/wafer. We divide these values to obtain 1.33 kWh/step as the fabrication energy of a "deposition" process step. Note that, the CNFETs and IGZO FETs in the M3D process are modeled at the 7 nm node, with feature sizes requiring EUV lithography. Thus, we leverage data reported for EUV lithography (in reference [4]) to determine the fabrication energy for each appropriate process step. A step-by-step guide that includes the exact calculations we performed to model the entire M3D process and all-Si process, is included in our Github repository.

To compute EPA for each process in Figure 2 (in kWh per 300 mm wafer), we multiply the two matrices shown in Equation 4. One matrix includes the number of times each process step is used in a process flow ($N_{step}^{(flow)}$ is the number of times process *step* is used in process *flow*). The other matrix includes the EPA of each process step. EPA is converted to grams CO$_2e$ using CI$_{\text{fab}}$, which is determined by the source of the electricity used by a foundry. Finally, we compute C$_{\text{embodied}}$ using Equation 2. The embodied carbon for the M3D IGZO/CNT/Si and all-Si processes is multiplied by $1.4\times$ to include a 40% overhead to account for the carbon footprint of the facility, as described in the 2015 International Technology Roadmap for Semiconductors [21]). Results are in Figure 2c.

$$\begin{bmatrix} \mathrm{N}^{\text{all-Si}}_{\text{FEOL}} & \mathrm{N}^{\text{M3D}}_{\text{FEOL}} \\ \mathrm{N}^{\text{all-Si}}_{\text{dry etch}} & \mathrm{N}^{\text{M3D}}_{\text{dry etch}} \\ \mathrm{N}^{\text{all-Si}}_{\text{EUV}} & \mathrm{N}^{\text{M3D}}_{\text{EUV}} \\ \mathrm{N}^{\text{all-Si}}_{\text{deposition}} & \mathrm{N}^{\text{M3D}}_{\text{deposition}} \\ \mathrm{N}^{\text{all-Si}}_{\text{wet etch}} & \mathrm{N}^{\text{M3D}}_{\text{wet etch}} \\ \mathrm{N}^{\text{all-Si}}_{\text{M/V 36 nm}} & \mathrm{N}^{\text{M3D}}_{\text{M/V 36 nm}} \\ \mathrm{N}^{\text{all-Si}}_{\text{M/V 48 nm}} & \mathrm{N}^{\text{M3D}}_{\text{M/V 48 nm}} \\ \mathrm{N}^{\text{all-Si}}_{\text{M/V 64 nm}} & \mathrm{N}^{\text{M3D}}_{\text{M/V 64 nm}} \\ \mathrm{N}^{\text{all-Si}}_{\text{M/V 80 nm}} & \mathrm{N}^{\text{M3D}}_{\text{M/V 80 nm}} \end{bmatrix}^T \begin{bmatrix} \mathrm{E}^{\text{iN7-EUV}}_{\text{FEOL}} \\ \mathrm{E}_{\text{dry etch}} \\ \mathrm{E}_{\text{EUV}} \\ \mathrm{E}_{\text{deposition}} \\ \mathrm{E}_{\text{wet etch}} \\ \mathrm{E}^{\text{EUV}}_{\text{M/V 36 nm}} \\ \mathrm{E}^{\text{SADP}}_{\text{M/V 42 nm}} \\ \mathrm{E}^{\text{ArFi LE2}}_{\text{M/V 64 nm}} \\ \mathrm{E}^{\text{ArFi LE}}_{\text{M/V 80 nm}} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 3 \\ 0 & 6 \\ 0 & 10 \\ 0 & 3 \\ 3 & 12 \\ 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}^T \begin{bmatrix} 436 \\ 2 \\ 3.\overline{3} \\ 1.\overline{3} \\ 0.\overline{3} \\ 37.74 \\ 32.85 \\ 23.07 \\ 19.57 \end{bmatrix} = \begin{bmatrix} 700.2 \\ 1080.2 \end{bmatrix} \tag{4}$$

III. Case Study: Cortex M0 + M3D-eDRAM

In this section, we demonstrate how to use our models to answer practical questions in carbon-efficient computing. As a case study, consider the following scenario. A design team is developing an embedded computing system to run a variety of applications that are well-represented by the workloads in Embench [17]. The team has decided on the target application lifetime (in months, including the expected use time per day), and high-level system architecture comprising a Cortex-M0 processor with on-chip program and data memories as shown in Figure 1. In deciding which technology to use to physically realize this computing system – considering trade-offs in its power, performance, area, and total carbon footprint – the team seeks to answer the following question: *Should it be implemented in the M3D process combining IGZO FETs + CNFETs + Si CMOS, or in the baseline Si process?*

A. Circuit-Level Schematics & Physical Design

We begin by introducing the schematic and physical layout of our embedded computing system (shown in Fig. 3), which we analyze using detailed simulations to quantify its Power consumption, Performance, Area, and total Carbon footprint (PPAtC), leveraging a combination of industry-standard EDA tools and our embodied carbon models from Section **II**-C. As described in Section **II**-A, our M3D-eDRAM is designed to combine the benefits of IGZO FETs, CNFETs, and Si FETs, while simultaneously overcoming their inherent challenges (summarized in Table I). To achieve the key characteristics listed below, we leverage a 3-transistor (3T) memory bit cell topology that comprises one IGZO FET and two CNFETs.

- *High memory density:* memory cells are integrated directly on top of peripheral circuits, reducing area footprint.
- *High data retention time:* retention time of the storage node (SN) is limited by ultra-low I_{OFF} of the IGZO FET.
- *High endurance:* eDRAM is charge-based, vs. devices that are not solid-state and exhibit relatively low endurance (e.g., RRAM [12]).
- *Low access latency:* read delay is limited by high CNFET I_{EFF}. Write delay is limited by high Si FET I_{EFF}.
- *Low access energy:* limited by dynamic energy of peripheral circuits and parasitic capacitance of wordlines/bitlines.
- *Low static power:* limited by peripheral circuits (DRAM cells do not consume static power, unlike SRAM cells).

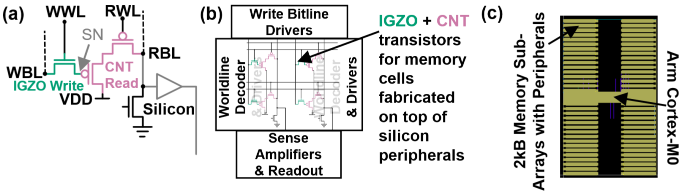
(a)
WWL
RWL
SN
RBL
WBL
IGZO Write
CNT
Read
VDD
Silicon
(b)
Write Bitline Drivers
Worldline Decoder
& Drivers
Sense Amplifiers & Readout
IGZO + CNT transistors for memory cells fabricated on top of silicon peripherals
(c)
2kB Memory Sub-Arrays with Peripherals
Arm Cortex-M0

Fig. 3. Circuit-level schematics & physical design. **(a)** 3T memory bit cell schematic. **(b)** eDRAM schematic including peripherals. **(c)** Chip floorplan.

B. Power, Performance, Area, and total Carbon

Here, we summarize our design flow for quantifying PPAtC of the embedded system running applications from Embench [17]. As described in the 5 steps below, we use a combination of cycle-accurate Register Transfer-Level (RTL) simulations (to count the number of eDRAM accesses), application-dependent power analysis after place-and-route (using Cadence Innovus), SPICE simulations of eDRAM circuit netlists (including wire parasitics), and compact models calibrated to experimental data for CNFETs, IGZO FETs, and Si CMOS.

Step 1) *Determine required memory size:* We compile each Embench application to run on the M0. We select the sizes (in kB) of the program & data memories so that the embedded system has sufficient memory to run any of the compiled applications. The program & data memories are each 64 kB.

Step 2) *eDRAM schematic & physical design:* We design our embedded system so that the M0 can read and write to the eDRAMs in a single clock cycle. Thus, in the design of both the M0 and the program/data memories, we must enforce that the critical path delay is shorter than the clock period (e.g., T_{CLK} = 2 ns for clock frequency f_{CLK} = 500 MHz). To facilitate fast critical path delay of the eDRAM (read/write access times), we partition the 64 kB into 2 kB sub-arrays, each with 512 32-bit words, which improves timing due to relatively smaller capacitive loading of 2 kB sub-arrays.

For eDRAM to meet timing, we adjust the following circuit design parameters in Figure 3: memory supply voltage (V_{DD} = 0.7 V to match the recommended V_{DD} for the ASAP7 standard cell libraries [19]), write word line voltage (V_{WWL} = 1.3 V to overdrive the IGZO FETs), and V_T of each FET (in the bit cells, write drivers, and sense amplifiers). For digital logic blocks (the decoder & refresh controller), we specify timing constraints in automated VLSI design flows. We validate timing using SPICE circuit simulations, with compact device models for Si CMOS [19], CNFETs [27], and IGZO FETs (using a virtual source model [37] with experimentally measured values: IGZO mobility = 1 cm^2/V.s and sub-threshold slope = 90 mV/decade for 44 nm gate length [38]).

Step 3) *M0 + eDRAM integration & physical design:* We export the memory layout (library exchange format: .lef), which can be instantiated together with the M0 core in standard VLSI digital design flows. We perform logic synthesis & place-and-route (Cadence Genus & Cadence Innovus) over a range of design parameters to generate multiple options for f_{CLK} and energy consumption of our system. In particular, we sweep the target clock frequency from 100 MHz to 1 GHz (in steps of 100 MHz), and sweep V_T of the FETs over all options offered in the ASAP7 standard cell library. Figure 4 shows the critical path delay for each design.

Step 4) *Application-dependent energy consumption:* We perform RTL simulations (using Synopsys VCS, with the program memory initialized with a compiled Embench application) to obtain cycle-accurate digital waveforms (digital 1 or 0 vs. time) for each net in our computing system. These waveforms are represented in .vcd format (value change dump), which enables us to: (a) determine the exact number of clock cycles to run an application, (b) determine the exact number of memory accesses and required data retention times (by analyzing reads/writes to specific memory addresses), and (c) determine the average energy per clock cycle for the M0 to run a specific application, as shown in Figure 4 (using Cadence Innovus to perform power analysis for a specific .vcd file).

Step 5) *Total carbon:* Equation 5 computes average $\text{C}_{\text{embodied}}$ *per good die* (considering yield) given our models from Section **II**. We use a die per wafer estimator [39] to find $N_{\text{diePerWafer}}$, considering physical die size (from place-and-route) and other relevant parameters (horizontal & vertical spacing of 0.1 mm, edge clearance of 5 mm, flat/notch height of 10 mm). As a demonstration, we consider 90% yield for our Si eDRAM and 50% yield for the M3D-eDRAM to reflect the relative maturity and complexity of each process. However, designers can choose arbitrary yield models (e.g., depending on technology node, process, and design robustness).

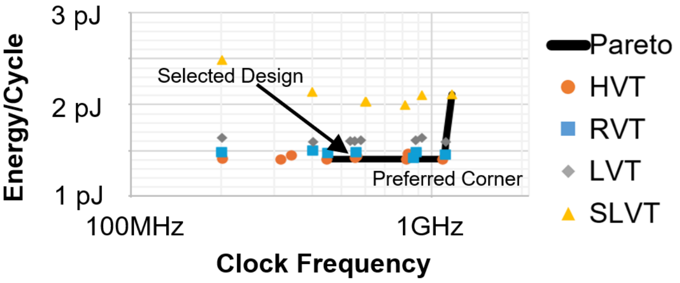

Fig. 4. Cortex-M0 average energy per cycle vs. clock frequency, shown for the *matmul-int* Embench workload. HVT, RVT, LVT, & SLVT for "High", "Regular", "Low" & "Super Low" V_T in the ASAP7 PDK [19].

$$C_{embodied}^{(good\ die)} = C_{embodied}^{wafer} \;/\; \left(N_{diePerWafer} \cdot Yield\right) \tag{5}$$

For $\text{C}_{\text{operational}}$, we consider running the "matmul-int" application for 2 hours per day for 24 months. Since $\text{CI}_{\text{use}}(t)$ varies throughout the day, we define an indicator function $\mathbb{I}_{8to10pm}(t)$, which is 1 from 8 pm to 10 pm, and 0 otherwise. We can then define power consumption vs. time according to Equation 6, where $P_{\text{static}}^{(\text{M0})}$ is the M0 static power, $E_{\text{dynamic}}^{(\text{M0})}$ is the M0 total dynamic energy to execute the application once, and $E_{\text{operational}}^{(\text{eDRAM})}$ is the eDRAM total operational energy to execute the application once (considering leakage power, eDRAM refresh, and application-specific read/write accesses).

$$P(t) = \left(P_{static}^{(M0)} + \frac{E_{dynamic}^{(M0)}}{N_{cycle} \cdot T_{clk}} + \frac{E_{operational}^{(eDRAM)}}{N_{cycle} \cdot T_{clk}} \right) \mathbb{I}_{8to10pm}(t) \tag{6}$$

We then substitute $P(t)$ into the general form for $\text{C}_{\text{operational}}$ (Equation 1), lumping all time-independent terms into $P_{operational}$ (Equation 7). Further reduction leads to Equation 8, where $\overline{\text{CI}}_{\text{use,8to10pm}}$ is the average value of CI_{use} from 8 pm to 10 pm over the entire lifetime.

$$C_{operational} = P_{operational} \int_0^{t_{life}} \text{CI}_{\text{use}}(t) \mathbb{I}_{8to10pm}(t) dt \tag{7}$$

$$C_{operational} = \overline{\text{CI}}_{\text{use,8to10pm}} \cdot P_{operational} \cdot t_{life} \cdot \frac{2 \text{ hours/day}}{24 \text{ hours/day}} \tag{8}$$

Table II summarizes PPAtC results from our design flow. In Section **III**-C, we quantify how these PPAtC trade-offs affect the *carbon efficiency* of each design.

TABLE II
PPAtC SUMMARY (CIRCUIT SIMULATION RESULTS ARE IN OUR GITHUB).

System	**M0 +** **Si eDRAM**	**M0 +** **IGZO/CNT/Si** **M3D-eDRAM**
clock frequency	500 MHz	500 MHz
M0 dynamic energy per cycle	1.42 pJ	1.42 pJ
average memory energy per cycle	18.0 pJ	15.5 pJ
clock cycles to run "matmul-int"	20,047,348	20,047,348
64 kB memory area footprint	0.068 mm^2	0.025 mm^2
total area footprint (memory + M0)	0.139 mm^2 H: 270 μm W: 515 μm	0.053 mm^2 H: 159 μm W: 334 μm
embodied carbon per wafer (U.S. grid)	837 kgCO$_2e$	1100 kgCO$_2e$
total die count per 300 mm wafer	299,127	606,238
embodied carbon per good die	3.11 gCO$_2e$	3.63 gCO$_2e$

C. Quantifying trade-offs in PPAtC

To quantify PPAtC trade-offs for the case study described at the beginning of Section **III**, we operate each computing system at the same clock frequency (f_{CLK} = 500 MHz), which corresponds to meeting a target latency constraint (i.e., each embedded application must finish executing in a fixed amount of time). To accurately quantify tC *per good die*, the area of each die (see Table II) and the number of good dies per wafer for each design must be calculated. Considering parameters such as number of dies per wafer and yield, the area per die of the all-Si design is 2.72× larger than the M3D design, but produces 1.13× more good dies per wafer. This translates into a 1.17× increase in $\text{C}_{\text{embodied}}$ *per good die* for the M3D design vs. the all-Si design (see Table II).

Figure 5 shows tC vs. lifetime, considering $\text{C}_{\text{embodied}}$ per good die, and $\text{C}_{\text{operational}}$ for 2 hours of use per day. In this case, $\text{C}_{\text{embodied}}$ dominates tC until 14 months for the all-Si design, and 19 months for the M3D design, after which $\text{C}_{\text{operational}}$ dominates. Before 11 months, tC is higher for the M3D design vs. the all-Si design, and then it switches to being higher for the all-Si design. In general, this crossover point depends on the application scenario and the design.

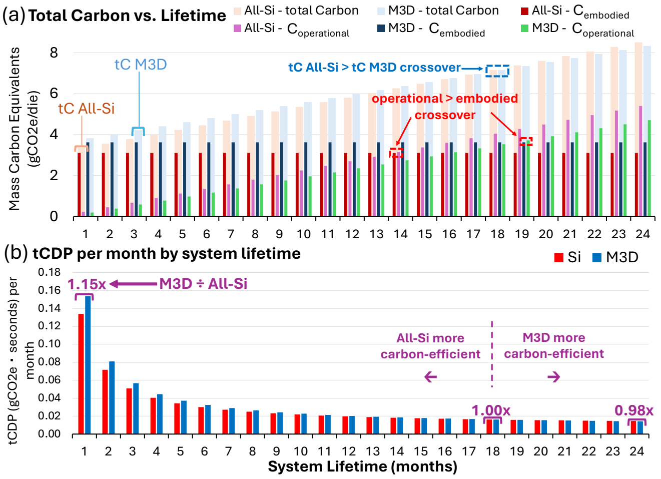

Fig. 5. **(a)** tC (gCO$_2e$) and **(b)** tCDP (gCO$_2e$/Hz) per month for all-Si and M3D designs vs. system lifetime (U.S. grid). tC is shown behind $\text{C}_{\text{embodied}}$ and $\text{C}_{\text{operational}}$ contributions. The ratio between M3D and all-Si tCDP is highlighted above 1, 18, and 24 months. It converges to the ratio of energy-delay product (EDP) as $\text{C}_{\text{operational}}$ dominates for long system lifetime.

D. tCDP

We compare the *carbon efficiency* of each design using the total carbon delay product metric (tCDP): the product of tC and application execution time, measured in gCO$_2e$/Hz [18]. This metric quantifies how efficiently CO$_2e$ emissions are being allocated to reach a certain clock frequency (an extensive discussion of the tCDP metric is in reference [18]). For a system lifetime of 24 months, the M3D design is 1.02× more carbon-efficient vs. the all-Si design (Fig. 5).

As described in the introduction, it is not clear whether M3D ICs are more *carbon-efficient* vs. Si CMOS ICs. Specifically, do the *energy efficiency benefits* of M3D ICs outweigh their $\text{C}_{\text{embodied}}$ drawback, considering factors such as area footprint, process complexity, and yield? We are now equipped to answer this question. As the answer depends on detailed analysis of several parameters, we use Figure 6a to visualize specific circumstances where M3D designs have better tCDP than all-Si designs. The y-axis corresponds to scaling the operational energy ($\text{E}_{\text{operational}}$) of the M3D design. For example, a value of 0.5 means that the $\text{E}_{\text{operational}}$ is 2.0× lower (leading to better tCDP). The x-axis corresponds to scaling $\text{C}_{\text{embodied}}$ of the M3D design. For example, a value of 2.0 means that $\text{C}_{\text{embodied}}$ is 2.0× higher (leading to worse tCDP). The colormap indicates the overall tCDP benefit of the M3D design vs. the all-Si design. The red shaded regions show when the M3D design is more carbon-efficient. Otherwise, the all-Si design is more carbon-efficient (blue shaded regions). The "tCDP isoline" represents the boundary between these two regions, which also depends on factors such as yield, system lifetime, and CI_{use}. The tCDP isoline is helpful in guiding designers to understand when it makes sense to pursue M3D technologies for carbon-efficient computing – considering a wide range of values for relative energy efficiency and $\text{C}_{\text{embodied}}$. Furthermore, Figure 6b shows how to expand this type of analysis to account for *uncertainty* in carbon accounting, e.g., *uncertainty* in yield, system lifetime, and CI_{use}. Uncertainty affects the position of the tCDP isoline. However, even in the presence of uncertainty, there are regions in which the M3D design maintains better tCDP vs. the all-Si design (and vice versa). This is especially useful information given the challenge of uncertainty in carbon accounting [18].

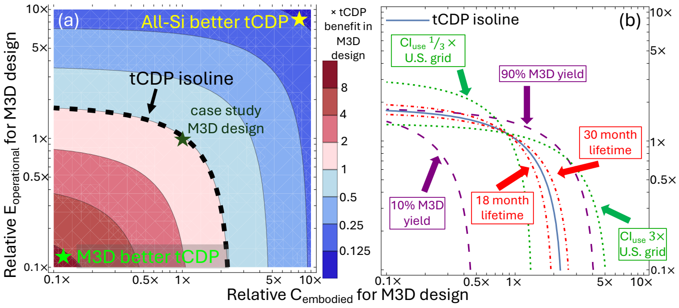

Fig. 6. **(a)** The colormap indicates the relative tCDP of the M3D design vs. the all-Si design. The dashed black line is the tCDP isoline. **(b)** Variation in the tCDP isoline due to: 6 month increase/decrease in system lifetime (red dashed lines), 3× increase/decrease in CI_{use} of the U.S. grid (green dashed lines), and 10%/90% M3D yield (purple dashed lines).

CONCLUSION

We demonstrate a detailed methodology for robustly quantifying trade-offs in power, performance, area, and total carbon footprint of future computing systems. We present a specific case study showing how to compare tCDP of an embedded system implemented in an M3D IGZO/CNFET/Si process vs. an all-Si process. This type of analysis can be extended to consider factors such as cost, new materials and processes, alternative memory cell topologies, water consumption, and more, to guide decision-making about which future technologies should be developed to reduce computing's carbon footprint.

ACKNOWLEDGMENTS

We acknowledge incredibly valuable support from the Salata Institute for Climate and Sustainability at Harvard, the Harvard Griffin Graduate School of Arts and Sciences Prize Fellowship, NSF Expeditions in Computing (NSF EiC: 2326605), NSF CIRC: 2346435, and NSF CAREER: 2443303.

References

[1] A. S. G. Andrae *et al.*, "On Global Electricity Usage of Communication Technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015. [Online]. Available: https://www.mdpi.com/2078-1547/6/1/117

[2] C. Freitag *et al.*, "The climate impact of ICT: A review of estimates, trends and regulations," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2102.02622

[3] S. Ayers *et al.*, "Measuring the Emissions and Energy Footprint of the ICT Sector: Implications for Climate Action," 2023. [Online]. Available: https://coilink.org/20.500.12592/hdr7xg6

[4] 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)*. [Online]. Available: https://doi.org/10.1109/IEDM13553.2020.9372004

[5] M. Elgamal *et al.*, "Carbon-Efficient Design Optimization for Computing Systems," in *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, ser. HotCarbon '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3604930.3605712

[6] 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*. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3470496.3527408

[7] 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. [Online]. Available: https://doi.org/10.1109/HPCA57654.2024.00058

[8] Z. Zhang *et al.*, "DeltaLCA: Comparative Life-Cycle Assessment for Electronics Design," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–29, 2024. [Online]. Available: https://doi.org/10.1145/3643561

[9] P. Raffeck *et al.*, "CO$_2$CoDe: Towards Carbon-Aware Hardware/Software Co-Design for Intermittently-Powered Embedded Systems," 2024. [Online]. Available: https://hotcarbon.org/assets/2024/pdf/hotcarbon24-final69.pdf

[10] G. Hills *et al.*, "Modern microprocessor built from complementary carbon nanotube transistors," *Nature*, vol. 572, no. 7771, pp. 595–602, 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-1493-8

[11] G. Fiori *et al.*, "Electronics based on two-dimensional materials," *Nature nanotechnology*, vol. 9, no. 10, pp. 768–779, 2014. [Online]. Available: https://doi.org/10.1038/nnano.2014.207

[12] H.-S. P. Wong *et al.*, "Metal–oxide RRAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012. [Online]. Available: https://doi.org/10.1109/JPROC.2012.2190369

[13] A. Belmonte *et al.*, "Tailoring IGZO-TFT architecture for capacitorless DRAM, demonstrating $> 10^3$ s retention, $> 10^{11}$ cycles endurance and Lg scalability down to 14nm," in *2021 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2021, pp. 10–6. [Online]. Available: https://doi.org/10.1109/IEDM19574.2021.9720596

[14] J. H. Lau, "Recent advances and trends in advanced packaging," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 12, no. 2, pp. 228–252, 2022. [Online]. Available: https://doi.org/10.1109/TCPMT.2022.3144461

[15] M. M. S. Aly *et al.*, "The N3XT approach to energy-efficient abundant-data computing," *Proceedings of the IEEE*, vol. 107, no. 1, pp. 19–48, 2018. [Online]. Available: https://doi.org/10.1109/JPROC.2018.2882603

[16] Y. Zhao *et al.*, "3D-Carbon: An Analytical Carbon Modeling Tool for 3D and 2.5D Integrated Circuits," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3649329.3658482

[17] "Embench™: Open Benchmarks for Embedded Platforms," 2021. [Online]. Available: https://github.com/embench/embench-iot

[18] M. Elgamal *et al.*, "CORDOBA: Carbon-efficient optimization framework for computing systems," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025. *To appear*.

[19] L. T. Clark *et al.*, "ASAP7: A 7-nm finFET predictive process design kit," *Microelectronics Journal*, vol. 53, pp. 105–115, 2016. [Online]. Available: https://doi.org/10.1016/j.mejo.2016.04.006

[20] "Electricity Maps," 2024. [Online]. Available: https://app.electricitymaps.com/map

[21] "International Technology Roadmap for Semiconductors 2.0 (2015 Edition): Environment, Safety, and Health," 2015. [Online]. Available: https://www.semiconductors.org/wp-content/uploads/2018/06/4_2015-ITRS-2.0-ESH.pdf

[22] J. D. Plummer *et al.*, *Integrated Circuit Fabrication: Science and Technology*. Cambridge University Press, 2023. [Online]. Available: https://doi.org/10.1017/9781009303606

[23] A. Belmonte *et al.*, "Lowest $I_{OFF} < 3\times\ 10^{-21}$ A/μm in capacitorless DRAM achieved by Reactive Ion Etch of IGZO-TFT," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2023, pp. 1–2.

[24] H. Hosono, "How we made the IGZO transistor," *Nature Electronics*, vol. 1, no. 7, pp. 428–428, 2018. [Online]. Available: https://doi.org/10.1038/s41928-018-0106-0

[25] Z. Pan *et al.*, "Approaches to Improve Mobility and Stability of IGZO TFTs: A Brief Review," *Transactions on Electrical and Electronic Materials*, pp. 1–9, 2024. [Online]. Available: https://doi.org/10.1007/s42341-024-00536-1

[26] G. Hills *et al.*, "Understanding energy efficiency benefits of carbon nanotube field-effect transistors for digital VLSI," *IEEE Transactions on Nanotechnology*, vol. 17, no. 6, pp. 1259–1269, 2018. [Online]. Available: https://doi.org/10.1109/TNANO.2018.2871841

[27] C.-S. Lee *et al.*, "A compact virtual-source model for carbon nanotube fets in the sub-10-nm regime—part i: Intrinsic elements," *IEEE transactions on electron devices*, vol. 62, no. 9, pp. 3061–3069, 2015. [Online]. Available: https://doi.org/10.1109/TED.2015.2457453

[28] J. Zhang *et al.*, "Carbon nanotube robust digital vlsi," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 4, pp. 453–471, 2012. [Online]. Available: https://doi.org/10.1109/TCAD.2012.2187527

[29] M. M. Shulaker *et al.*, "Efficient metallic carbon nanotube removal for highly-scaled technologies," in *2015 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2015, pp. 32–4. [Online]. Available: https://doi.org/10.1109/IEDM.2015.7409815

[30] S. B. Boyd, *Life-cycle assessment of semiconductors*. Springer Science & Business Media, 2011. [Online]. Available: https://doi.org/10.1007/978-1-4419-9988-7

[31] H. Y. Teah *et al.*, "Life cycle greenhouse gas emissions of long and pure carbon nanotubes synthesized via on-substrate and fluidized-bed chemical vapor deposition," *ACS Sustainable Chemistry & Engineering*, vol. 8, no. 4, pp. 1730–1740, 2020. [Online]. Available: https://doi.org/10.1021/acssuschemeng.9b04542

[32] R. S. Park *et al.*, "Hysteresis-free carbon nanotube field-effect transistors," *ACS nano*, vol. 11, no. 5, pp. 4785–4791, 2017. [Online]. Available: http://dx.doi.org/10.1021/acsnano.7b01164

[33] Y. Su *et al.*, "Monolithic 3-D Integration of Counteractive Coupling IGZO/CNT Hybrid 2T0C DRAM and Analog RRAM-Based Computing-In-Memory," *IEEE Transactions on Electron Devices*, 2024. [Online]. Available: https://doi.org/10.1109/TED.2024.3372937

[34] J. K. Lee *et al.*, "Self-Aligned Top-Gate IGZO TFT with Stepped Structure for Suppressing Short Channel Effect," *IEEE Electron Device Letters*, 2023. [Online]. Available: https://doi.org/10.1109/LED.2023.3317403

[35] Y. Guan *et al.*, "Ultra-thin top-gate insulator of atomic-layer-deposited HfO$_x$ for amorphous InGaZnO thin-film transistors," *Applied Surface Science*, vol. 625, p. 157177, 2023. [Online]. Available: https://doi.org/10.1016/j.apsusc.2023.157177

[36] Y. Zhang *et al.*, "Sub-100 nm self-aligned top-gate amorphous InGaZnO thin-film transistors with gate insulator of 4 nm atomic-layer-deposited AlO$_x$," *IEEE Electron Device Letters*, vol. 44, no. 3, pp. 444–447, 2023. [Online]. Available: https://doi.org/10.1109/LED.2023.3237747

[37] A. Khakifirooz *et al.*, "A simple semiempirical short-channel MOSFET current–voltage model continuous across all regions of operation and employing only physical parameters," *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1674–1680, 2009. [Online]. Available: https://doi.org/10.1109/TED.2009.2024022

[38] S. Samanta *et al.*, "Amorphous IGZO TFTs Featuring Extremely-Scaled Channel Thickness and 38 nm Channel Length: Achieving Record High Gm,max of 125 μS/μm at VDS of 1 V and ION of 350 μA/μm," in *2020 IEEE Symposium on VLSI Technology*, 2020, pp. 1–2. [Online]. Available: https://doi.org/10.1109/VLSITechnology18217.2020.9265052

[39] "Die Per Wafer Estimator," 2024. [Online]. Available: https://anysilicon.com/die-per-wafer-formula-free-calculators/