Reducing Smart Phone Environmental Footprints with In-Memory Processing

Zhuoping Yang
School of Engineering
Brown University
Providence, USA
zhuoping_yang@brown.edu

Wei Zhang
School of Engineering
Brown University
Providence, USA
wei_zhang6@brown.edu

Shixin Ji
School of Engineering
Brown University
Providence, USA
shixin_ji@brown.edu

Peipei Zhou
School of Engineering
Brown University
Providence, USA
peipei_zhou@brown.edu

Alex K. Jones
Department of EECS
Syracuse University
Syracuse, USA
akj@syr.edu

***Abstract*—Smart phones have revolutionized the availability of computing to the consumer. Recently, smart phones have been aggressively integrating artificial intelligence (AI) capabilities into their devices. The custom designed processors for the latest phones integrate incredibly capable and energy efficient graphics processors (GPUs) and tensor processors (TPUs) to accommodate this emerging AI workload and on-device inference. Unfortunately, smart phones are far from sustainable and have a substantial carbon footprint that continues to be dominated by environmental impacts from their manufacture and far less so by the energy required to power their operation. In this paper we explore the possibility of reversing the trend to increase the dedicated silicon dedicated to emerging application workloads in the phone. Instead we consider how in-memory processing using the DRAM already present in the phone could be used in place of dedicated GPU/TPU devices for AI inference. We explore the potential savings in embodied carbon that could be possible with this tradeoff and provide some analysis of the potential of in-memory computing to compete with these accelerators. While it may not be possible to achieve the same throughput, we suggest that the responsiveness to the user may be sufficient using in-memory computing, while both the embodied and operational carbon footprints could be improved. Our approach can save circa 10–$15\,\mathrm{kg\,CO_2e}$.**

***Index Terms*—sustainability, embodied carbon, in-memory processing, AI, inference**

I. Introduction

Smart phones have changed the landscape of how computing is used in society. Since their introduction, more and more tasks have become mobile friendly to the point where some tasks such as ticketing for air travel, map directions, photography, and access to social networks have become more difficult on traditional devices compared to their mobile counterparts. Furthermore, generation after generation, these devices continue to be called upon for increasingly complex tasks.

This work is supported in part by NSF awards #2213701, #2217003, #2324864, #2328972.

An enabling technology to allow this dramatic increase in capability from smart phones has been dark silicon and custom hardware acceleration. For instance, incredibly compute intensive algorithms like signal processing for the wireless radio, video and auto decompression, and encryption ciphers have been integrated into the processor with custom hardware blocks. This sort of dark silicon optimization targeting energy efficiency to satisfy the needs of increasing numbers of complex tasks has made new capabilities possible while maintaining or increasing the operational lifetime on a single battery charge. The emergence of artificial intelligence (AI) as a consumer facing tool has recently pushed these devices to further increase dark silicon. Processors customized for the latest phones include embedded graphics processors (GPUs) and tensor processors (TPUs) to efficiently compute AI inference tasks.

Unfortunately, dark silicon has had an unintended side effect. The trends to grow silicon area in spite of technology scaling to increasing small feature sizes increases the semiconductor contributions to embodied carbon in these devices [1]. Embodied carbon is more balanced with operational carbon in traditional and server class machines [2], [3]. Moreover, the embodied carbon is dominated by memory and storage such as the many DRAM chips in DIMMs as well as the Flash chips for solid-state storage in server class machines [3]–[6]. Smart phones are much more heavily dominated by embodied carbon from chips for processing.

Thus, exploring alternate methods to compute emerging applications efficiently with a smaller silicon area will go a long way to reducing the carbon footprint of these devices. Considering there are more mobile phones (circa 8.6 billion) than people (circa 7.9 billion) as of 2022, this platform has a huge impact on world sustainability.

Processing in DRAM has been proposed for commodity DRAM with minimal [7]–[9] to no fundamental changes [10]–[13] to commercial devices. We explore the most recent pro-

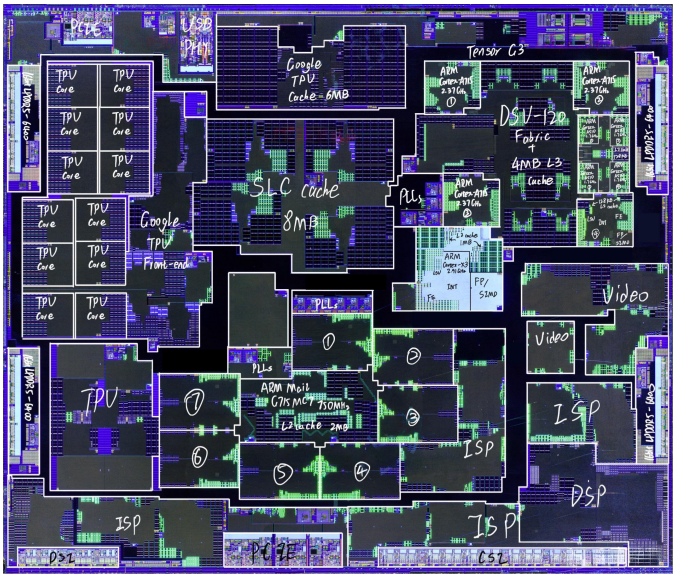


Fig. 1: Chip Layout for Google Tensor G3 [14].

posals to use processing in commodity DRAM to explore the potential of eliminating the need for expensive GPU and TPU devices to compute these expensive workloads (particularly AI inference) and the impact to sustainability of these phones.

In this paper we explore the novel use of in-memory processing as a *sustainable replacement* for using dedicated silicon area for GPU and TPU units and potentially other dedicated hardware accelerators. The area of processors designed for smart phones continues to increase in spite of continued scaling, most recently to 4 nm and ultimately 3 nm in the latest processors. We show that the most recent DRAM processing in memory can improve the throughput per area (performance per embodied CO_2) and the throughput per energy (performance per operational CO_2).

In the next section we provide more details on the overhead of dark silicon in modern smart phones.

II. Progression of Smart Phone Processors

The amount of dark silicon has been increasing with each new smart phone generation [1]. In Fig. 1 we show an annotated layout for the Google Tensor G3 chip [14] from the Pixel 8 smart phone. The entire left side of the chip is dedicated to the TPU, with a TPU cache in top left portion of the chip. The bottom middle portion of the chip is dominated by the Arm Mali-G715 GPU, which encircles its own GPU cache. There is some dedicated hardware associated with processing IO at the chips extremities. Finally the main core processor is at the top right containing nine cores, an Arm Cortex-X3 core with floating-point and vector (SIMD) units, and four each of Cortex-A715 (big cores) and Cortex-A510 (little cores) surrounding 4MB L3 cache. In the center of the chip is a system-level cache of 8MB. By estimating just the portion of the chip dedicated for the GPU and TPU, this requires approximately 50% of the chip area.

In Fig. 2 we show a similar layout for the A17 Pro chip that powers the iPhone15 line of phones. The GPU takes most

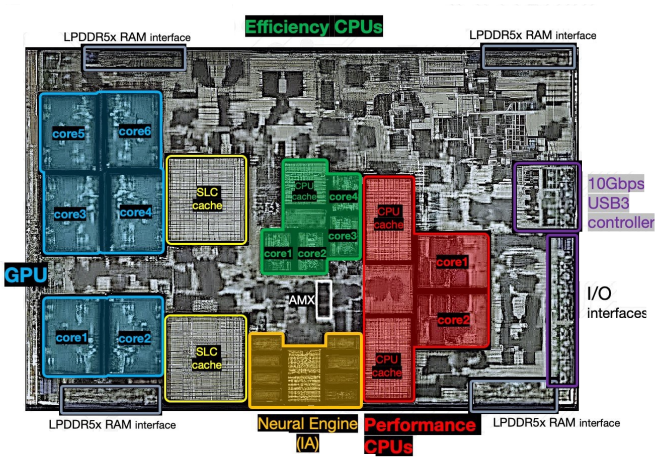


Fig. 2: Chip Layout for Apple A17 Pro [15].

of the left quarter of the chip, while the tensor processor is at the bottom, with system level cache between them. Like the google processor, there is big/little layout with two high performance cores with substantial cache real-estate. There are four little cores with a shared cache. All told the GPU and TPU take about a third of the chip, while the CPUs, even in their big/little format, take less than a quarter. The remainder of the space is dedicated for unknown dark silicon.

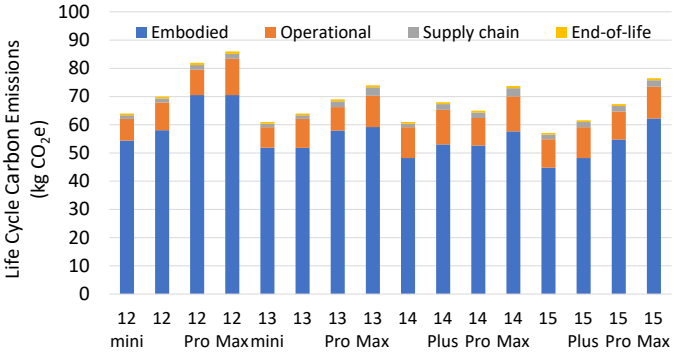


Fig. 3: Life cycle carbon emissions of Apple iPhones from iPhone 12 to iPhone 15 series [16].

Prior work has identified that embodied carbon of smart phones has dominated their carbon footprint [1], [17], [18]. Embodied silicon can be as much as 90% of the embodied carbon and the trends do not seem to be improving this factor. For instance, the iPhone carbon foot print calculations for the last four generations are shown in Fig. 3. The iPhone 12 released in 2020 has a substantially similar footprint to the iPhone 15 which is current as of this writing. From iPhone 12 mini to iPhone 12 Pro Max, the embodied carbon is circa 55–70 kg CO_2e[1]. For the iPhone 15 generation this remains at 45–62 kg CO_2e.

The Google Pixel phones have a similar trend such that the embodied carbon of the Google Pixel 5 from circa 2020 ranged from 45–67 kg CO_2e (noting there was no "Pro" model for this generation and the Pixel 6 Pro reached almost 80 kg CO_2e)

[1]Greenhouse gas emissions are measured in weight of CO_2 equivalent.

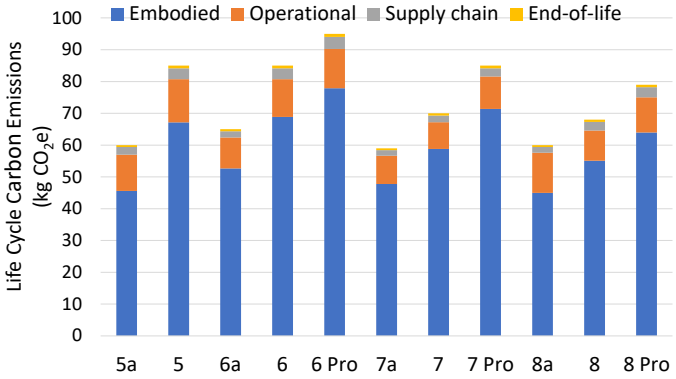


Fig. 4: Life cycle carbon emissions of Google Pixel Phones from Pixel 5 to Pixel 8 series [19].

TABLE I: Statistics and trends for Google, Apple smart phone procesors. *Reported 4 nm but similar to 5 nm. **estimated.

Name	Size mm^2	Trans.	Year	Fab Size	Models
Snap 765G	84	3B	2019	7 nm	4a, 5a, 5 Pro
A14 Bionic	88	8.5B	2020	5 nm	12, Mini, Pro, Max
Tensor	108	unknown	2021	5 nm	6a, 6, 6 Pro
A15 Bionic	108	15B	2021	5 nm	13, Mini, Pro, Max, 14, Plus
Tensor G2	114	15.8B	2022	5 nm	7a, 7, 7 Pro
A16 Bionic	113	16B	2022	4 nm*	14, Pro, Max, 15, 15 Plus
Tensor G3	135	unknown	2023	4 nm	8a, 8, 8 Pro
A17 Pro	103	19B	2023	3 nm	15 Pro, Max

and that the most recent line of Pixel 8 devices ranged from 45–63 kg CO_2e.

These trends seem to indicate that embodied carbon *is* decreasing slightly across the different generations, which is a welcome sign. Unfortunately, the trends in the semiconductor devices are that both the processes are becoming more expensive from a carbon perspective [20], [21] and that the area of the semiconductors are *still increasing* [1]. Table I shows the statistics of the processors for these phone generations [22]. These results show, if anything, the opposite trend compared to the embodied carbon reductions. Transistor count has more than doubled in the last four years, and the die size is monotonically increasing even as the fabrication technology feature size continues to descend.

Because the embodied carbon of the smart phone includes items beyond the chips, it is expected that these savings are coming from the manufacturing of other devices such as the screen, battery, etc. There remains an opportunity to make strides in improving embodied carbon in smart phones by finding creative ways to reduce the silicon real-estate.

One method to accomplish this is to reduce the number of dedicated dark silicon accelerators and to better identify those accelerators that are absolutely essential [1]. However, from the die charts from Figs. 1 and 2 indicate that alternative solutions to accelerating AI tasks from using the tensor and graphics processing accelerator can significant reduce chip area. We explore this in the next section.

III. Sustainable Inference In-Memory

Numerous proposals demonstrate in-DRAM computing using charge-sharing [7]–[9] or by intentionally violating memory timing parameters [12], [26]. These approaches leverage the use of DRAM micro instructions for commodity DRAM with minimal [7]–[9] to no fundamental changes [10]–[13] on commercial devices. We leverage the SIMDRAM [9] and Count2Multiply [27] approaches to implement tensor kernels from large language models (LLMs) LLaMA [23] and LLaMA [24] shown in Table II [27]. Our DRAM parameters are relayed in Table III. We normalized the results to a NVIDIA RTX 3090 GPU implementation.

TABLE II: GEM**V** and GEM**M** dimensions from [23], [24]

Model	ID	M	N	K	ID	M	N	K
LLaMA	V0	1	22016	8192	M0	8192	22016	8192
LLaMA	V1	1	8192	22016	M1	8192	8192	22016
LLaMA-2	V2	1	8192	8192	M2	8192	8192	8192
LLaMA-2	V3	1	28672	8192	M3	8192	28672	8192
LLaMA-2	V4	1	8192	28672	M4	8192	8192	28672

TABLE III: Memory organization and architectural parameters

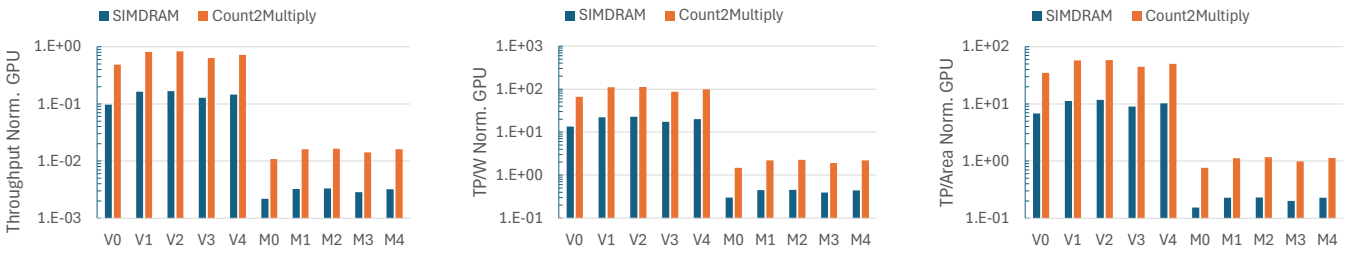
DRAM	Memory Controller 8 kB row size, FR-FCFS scheduling Main Memory DDR4-2400, 1 channel, 1 rank, 8 devices + ECC DRAM chip 4 banks, 1 kB row size, 1024 rows per subarray

Fig. 5 presents throughput and throughput per Watt and area normalized to the RTX 3090 GPU baseline. The in-memory accelerators cannot meet the performance of the server class GPU (Fig. 5a). However, the throughput per energy (Fig. 5b) and throughput per area (Fig. 5c), SIMDRAM configurations still provide competitive results while Count2Multiply consistently outperforms the GPU. While this comparison is against a server class GPU. So called "embedded" GPUs tend to be smaller and have fewer SIMD arrays to keep their device power down. However, the architecture of their SIMD arrays is fundamentally similar to their server class counterparts. Thus, their performance per Watt and area tends to be consistent [28], which makes these results similarly consistent.

Based on the in-memory capability we explored the savings from removing the GPU and TPU from the processor footprint of the smart phone. We this savings in three scenarios, that the processor contributes 50%, 30%, and 20% of the total embodied carbon to the phone. We show the new carbon footprint for Google phones in Fig. 6 and the same for Apple phones in Fig. 7. Based on these results we determine that this approach can reduce embodied carbon between 11.3 kg CO_2e and 19.5 kg CO_2e with an average savings of 14.9 kg CO_2e for Google phones and saving as much as 25% of the embodied carbon footprint. For apple devices the values range from 7.4 kg CO_2e to 11.6 kg CO_2e with an average savings of 9.2 kg CO_2e, saving as much as 17% of the embodied carbon. To put these results into perspective, the Apple environmental reports indicate that 128G of Flash solid-state storage is 9 kg CO_2e, making these savings quite valuable.

IV. Conclusions and Future Work

Embodied carbon is a significant source of carbon in modern smart phones. In-memory computing can provide efficient alternatives to GPU and TPU processors added into modern smart phone processors. Our results demonstrate that for LLMs, it may be possible to save between 7.4 and



(a) Throughput (b) Performance per Watt (c) Performance per Area

Fig. 5: GPU-normalized performance for ternary GEMM and GEMV [23]–[25].

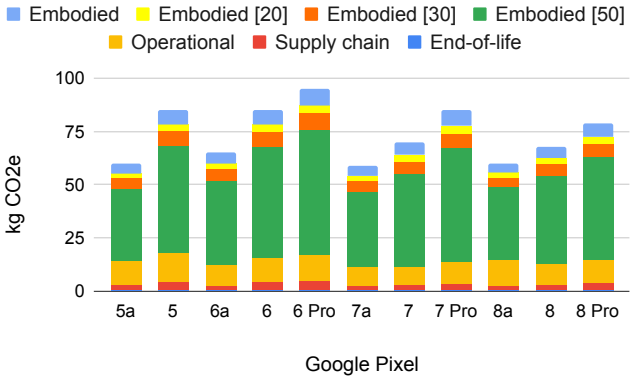


Fig. 6: Embodied carbon reduction for Google phones.

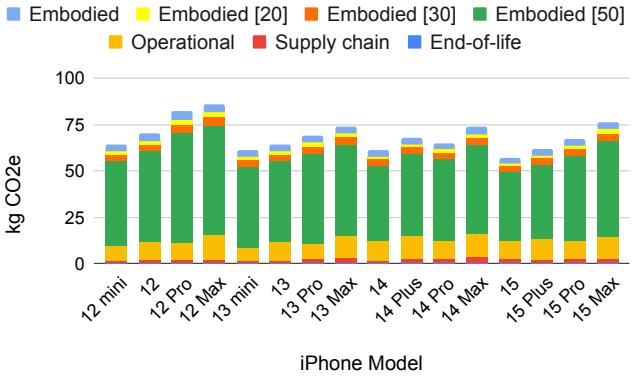


Fig. 7: Embodied carbon reduction for Apple phones.

14.9 kg CO_2e which is equivalent to the carbon cost of 106 to 212 GB of Flash. In the future we will investigate whether there are hardware blocks that can be eliminated combined with other sources of embodied carbon savings.

References

[1] E. Brunvand *et al.*, "Dark Silicon Considered Harmful: A Case for Truly Green Computing," in *IGSC*, 2018, pp. 1–8.

[2] S. Ji *et al.*, "SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators," in *ISVLSI*, 2024, pp. 1–6.

[3] "Life Cycle Assessment of Dell R740," [Available Online] https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/Full_LCA_Dell_R740.pdf, 2019.

[4] D. Berger *et al.*, "Research avenues towards net-zero cloud platforms," Azure Systems Research, February 2023, [Available Online] https://netzero.sysnet.ucsd.edu/pdf/netzero23-berger.pdf.

[5] J. Wang *et al.*, "Designing cloud servers for lower carbon," in *ISCA*, 2024, pp. 452–470.

[6] I. Samaye *et al.*, "Towards Sustainable Low Carbon Emission Mini Data Centres," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2405.01909

[7] V. Seshadri *et al.*, "Ambit: In-memory accelerator for bulk bitwise operations using commodity dram technology," in *MICRO*, 2017.

[8] X. Xin *et al.*, "Elp2im: Efficient and low power bitwise operation processing in dram," in *HPCA*. IEEE, 2020, pp. 303–314.

[9] N. Hajinazar *et al.*, "Simdram: a framework for bit-serial simd processing using dram," in *ASPLOS*, 2021, pp. 329–345.

[10] V. Seshadri *et al.*, "Rowclone: Fast and energy-efficient in-dram bulk data copy and initialization," in *MICRO*, 2013, pp. 185–197.

[11] S. Roy *et al.*, "Pim-dram: Accelerating machine learning workloads using processing in commodity dram," *JETCAS*, vol. 11, no. 4, pp. 701–710, 2021.

[12] I. E. Yuksel *et al.*, "Functionally-complete boolean logic in real dram chips: Experimental characterization and analysis," in *HPCA*, 2024.

[13] Y. Paik *et al.*, "Achieving the performance of all-bank in-dram pim with standard memory interface: Memory-computation decoupling," *IEEE Access*, vol. 10, pp. 93 256–93 272, 2022.

[14] "Quadrans Muralis on X: "Google Tensor G3 die shot. Die size = approx. 135.2 mm^2," [Available Online] https://x.com/QaM_Section31/status/1797279392374411340, 2024.

[15] "High Yield on X: "Probably the best Apple A17 Pro die-shot analysis yet, especially with such a high-res picture." / X," [Available Online] https://x.com/highyieldYT/status/1711453511848706228, 2023.

[16] "Environment - apple," [Available Online] https://www.apple.com/environment/, 2024.

[17] D. Kline *et al.*, "Sustainable ic design and fabrication," in *IGSC*, 2017.

[18] U. Gupta *et al.*, "ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool," in *ISCA*, 2022, pp. 784–799.

[19] "Sustainability reports & case studies - google sustainability," [Available Online] https://sustainability.google/reports/, 2024.

[20] A. K. Jones *et al.*, "Considering fabrication in sustainable computing," in *ICCAD*, 2013, pp. 206–210.

[21] M. Garcia Bardon *et al.*, "DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies," in *2020 IEDM*.

[22] "Google Pixel 6a vs. Pixel 6 vs. Pixel 6 Pro — CNN Underscored," [Available Online] https://www.cnn.com/cnn-underscored/electronics/google-pixel-6a-vs-pixel-6-vs-pixel-6-pro, 2023.

[23] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[24] ——, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[25] S. Ma *et al.*, "The era of 1-bit llms: All large language models are in 1.58 bits," *arXiv preprint arXiv:2402.17764*, 2024.

[26] F. Gao *et al.*, "Computedram: In-memory compute using off-the-shelf drams," in *MICRO*, 2019, pp. 100–113.

[27] J. P. C. de Lima *et al.*, "Count2multiply: Reliable in-memory high-radix counting," arXiv preprint arXiv:2409.10136, 2024.

[28] P. Dong *et al.*, "EQ-ViT: Algorithm-Hardware Co-Design for End-to-End Acceleration of Real-Time Vision Transformer Inference on Versal ACAP Architecture," *IEEE TCAD*, 2024.