Characterizing the Intergenerational Carbon Footprint of AI Infrastructure

Ruben Verheyden Lieven Eeckhout
Ghent University, Belgium

***Abstract*—This paper proposes a framework for characterizing the carbon footprint of AI infrastructure across multiple generations of hardware accelerators. While relying on carbon data for three recent generations of Google TPU accelerators, we find that AI model complexity growth largely outpaces AI hardware carbon efficiency improvements which leads to a significant and rapid increase in the cumulative and instantaneous total carbon footprint of modern-day AI infrastructure.**

I. Introduction

Artificial intelligence (AI) is a critical driver for modern-day technological innovations across many domains including healthcare, manufacturing, industrial operations, engineering, and science. However, the AI revolution comes at a significant environmental cost. The fast-paced growth of AI is putting increasing pressure on specialized hardware and high-energy data centers [11, 12]. Recent studies confirm this trend. Indeed, the Information and Communication Technology (ICT) sector as a whole contributes to 2.1% to 3.9% of global carbon emissions, surpassing the aviation and shipping industry [4]. The International Energy Agency (IEA) reports that in 2024, datacenters consumed about 415 terawatt-hours (TWh) of electricity, or roughly 1.5% of global electricity, while projecting that it will more than double to 945 TWh by 2030 [2]. Similarly, the US Department of Energy reports that by 2028, datacenter computing will account for 6.7% to 12% of the total electricity consumption in the US [10].

Recognizing the growing importance of sustainable operation, companies release life-cycle assessments (LCA) reports for their products [5]. More specifically for AI, Schneider et al. [9] provide an LCA for three generations of Google's specialized AI Tensor Processing Unit (TPU) accelerator. While LCA reports are extremely valuable to assess the environmental footprint for individual devices, there are at least two major limitations when looking at the broader picture.

First, an LCA report lacks an intergenerational perspective. Indeed, it looks at a device in isolation which can be misleading given the rapid succession of AI hardware accelerators, e.g., Google released TPU v5 in 2023, v6 in 2024, and v7 in 2025 [3, 9]. Older devices are not immediately replaced and continue to operate alongside newer generations, i.e., the newest generation is used to run the most recent AI model while previous generations are used to continue to run older models. This limitation has been partially addressed by Plotnik et al. [8] who analyze how embodied carbon emissions accumulate across hardware generations.

Second, an LCA report lacks a scale-out perspective, i.e., it lacks a perspective on how many devices are deployed at a given point in time. Indeed, AI model complexity is growing at a fast pace, i.e., at a rate of 4.5× per year according to EpochAI [3]. This requires an increasingly large number of accelerators to be deployed for supporting the ever growing AI models. For example, OpenAI and Nvidia recently announced to "*build and deploy at least 10 GW of AI datacenters with Nvidia systems representing millions of GPUs for OpenAI's next-generation AI infrastructure*" [6].

In this paper, we propose a framework to reason about the intergenerational carbon footprint of AI infrastructure while considering the emissions due to manufacturing (i.e., embodied footprint) and device use throughout its lifetime (i.e., operational emissions). We compute (1) the *cumulative total carbon* to characterize the grand total carbon footprint across coexisting generations of AI infrastructure, and (2) the *instantaneous total carbon* to quantify AI infrastructure's carbon emissions at a given moment in time. We use this framework to analyze the carbon footprint across three generations of Google's TPU (v4i, v5e and v6e) and conclude that even though individual TPUs have become increasingly carbon-efficient (i.e., by 2.93× from v4i to v6e [9]), the cumulative total carbon increases by 5.98× and the instantaneous total carbon increases by 1.44× from v4i to v6e, when assuming constant AI model complexity. However, when considering AI model complexity growth, the cumulative total carbon increases by 531× and the instantaneous total carbon increases by 359×. This rapid growth in total carbon is a result of AI model growth largely outpacing the carbon efficiency improvements of AI hardware accelerators.

II. Intergenerational Carbon Footprint

We now present a framework to quantify the carbon footprint of an AI hardware family across successive generations. It explicitly accounts for the coexistence of multiple hardware generations, and it models both embodied and operational emissions at two levels: cumulative total and instantaneous (per unit of time) total carbon emissions.

A. Cumulative Total Carbon

We first compute the cumulative total carbon footprint of AI infrastructure, which quantifies the grand total footprint across multiple generations. The *cumulative total carbon* (with

superscript c) is the sum of the cumulative embodied and operational carbon:

$$C^c_{total} = C^c_{emb} + C^c_{op}. \quad (1)$$

The cumulative embodied footprint is computed as the sum of the embodied footprint across successive generations i:

$$C^c_{emb} = \sum_i C_{emb,i}. \quad (2)$$

Likewise, the cumulative operational footprint is computed as the sum of the operational footprint across generations:

$$C^c_{op} = \sum_i C_{op,i}. \quad (3)$$

The operational footprint of generation i is computed as the *operational carbon rate* $C^r_{op,i}$ multiplied with the *time in use* t_i, i.e., the time since deployment of this generation, or the product of power consumption P_i with the time in use t_i and carbon intensity *CI*:

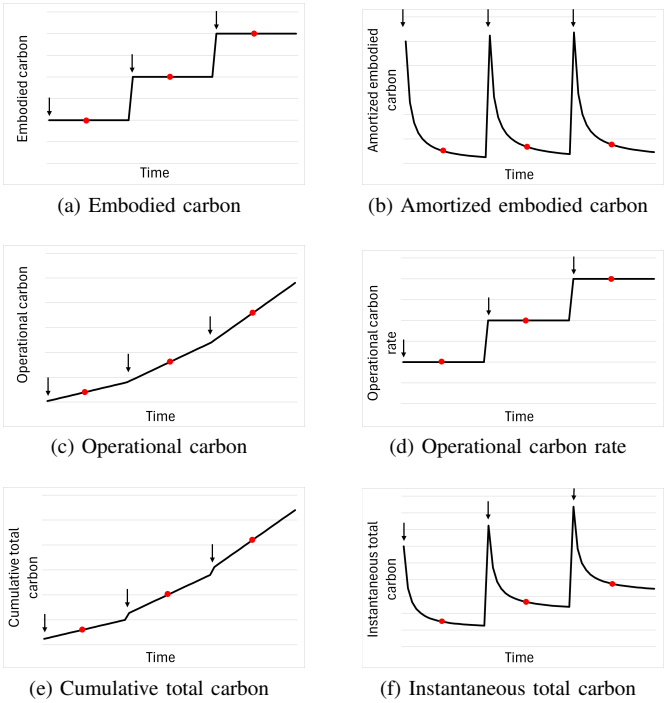
$$C_{op,i} = C^r_{op,i} \times t_i = P_i \times t_i \times CI. \quad (4)$$

Carbon intensity is measured in gram CO_2 emissions per unit of energy (kilowatt-hours) and depends on the location of AI infrastructure deployment. The world's average carbon intensity equals 472 gCO_2/kWh while varying substantially across different geographical locations depending on the mix of brown (coal, gas) and green (wind, solar, hydropower, nuclear) energy sources, e.g., India (707 gCO_2/kWh), China (555 gCO_2/kWh), United States (384 gCO_2/kWh), Europe (287 gCO_2/kWh), Norway (29 gCO_2/kWh) [1].

Figure 1 (left column) illustrates how the cumulative total carbon footprint evolves over time as new generations are added to the AI infrastructure — the deployment of new generations is indicated by the vertical arrows. The embodied footprint follows a staircase curve as each new generation adds a fixed embodied carbon footprint upon deployment, see Figure 1(a). The operational footprint follows a piecewise linear curve, see Figure 1(c): each generation incurs a different operational carbon rate which, when multiplied with the time in use, leads to a linearly increasing cumulative operational footprint. Note that the slope increases due to the cumulative operational footprint of successive generations. Adding the embodied and operational footprint leads to the curve shown in Figure 1(e): the deployment of a new generation incurs a step in the total carbon footprint (i.e., adding embodied footprint), which is followed by an increasingly steeper linear curve due to the use of the new generation (i.e., increasingly higher cumulative operational carbon rate).

B. Instantaneous Carbon Footprint

While the total cumulative footprint is useful for understanding AI infrastructure's grand total carbon footprint since its initial deployment, in practice, it may be more useful to compute AI infrastructure's instantaneous total carbon, i.e., the total footprint of the AI infrastructure at a given point in time. This enables analyzing how the AI infrastructure's carbon emissions evolve over time.



(a) Embodied carbon (b) Amortized embodied carbon

(c) Operational carbon (d) Operational carbon rate

(e) Cumulative total carbon (f) Instantaneous total carbon

Fig. 1: Illustration of the cumulative and instantaneous total carbon footprint across different generations of AI hardware.

To do so, we divide the embodied and operational footprint by the time in use. More specifically, for the embodied footprint, this leads to the *amortized embodied carbon*:

$$C^a_{emb} = \sum_i \frac{C_{emb,i}}{t_i}, \quad (5)$$

which spreads the "sunk" embodied carbon of a generation over time. The amortized embodied carbon is an instantaneous metric that quantifies the embodied carbon at a given moment in time per unit of time, e.g., per month as done in this paper.[1] For example, if a new generation of AI infrastructure incurs a total embodied carbon footprint of 1,000 $kgCO_2$, its instantaneous amortized embodied carbon per month equals $1{,}000/n$ $kgCO_2$ after n months since deployment.

Dividing the total operational footprint by the time in use leads to the *operational carbon rate*:

$$C^r_{op} = \sum_i C^r_{op,i} = \sum_i \frac{C_{op,i}}{t_i} \quad (6)$$

which effectively computes the instantaneous operational carbon footprint per month, or in other words, the total carbon emissions per month for operating the AI infrastructure.

Adding the amortized embodied footprint and the operational footprint rate leads to the *instantaneous total carbon*

[1]The unit of time could be any notion of a time period, e.g., a week, month, quarter, year. Without loss of generality, we consider a month in this paper.

	TPU v4i	TPU v5e	TPU v6e
Date	Jan 2020	Aug 2023	May 2024
Embodied carbon footprint EC_i ($kgCO_2$)	386	402	692
Operational carbon rate OC_i^r ($kgCO_2$ per month)	43.57	43.11	79.99
Normalized performance p_i	1×	1.16×	5.36×
CCI (gCO_2/ExaFLOP)	1,011	863	345
Carbon efficiency	1×	1.17×	2.93×
Training compute (ZettaFLOP)	40	8,760	27,080
Normalized operation count O_i	1×	219×	677×

TABLE I: Three generations of Google's versatile TPU [9] (top part of the table), alongside AI model training complexity [3] (bottom part of the table).

which quantifies AI infrastructure's total footprint per month across multiple AI hardware generations (note superscript i):

$$C_{total}^{i} = C_{emb}^{a} + C_{op}^{r}. \tag{7}$$

This is illustrated in Figure 1 (right column). The amortized embodied footprint follows a hyperbolically decreasing curve with each new generation of hardware, see Figure 1(b). Note that by the time a new generation is added, the previous generation's embodied footprint may not have been fully amortized. The operational footprint rate follows a staircase curve with each step denoting the increase in operational carbon rate with each generation added to the AI infrastructure, see Figure 1(d). Adding the amortized embodied footprint and the operational footprint rate curves leads to the instantaneous total carbon curve shown in Figure 1(f). The introduction of a new generation of hardware leads to a spike (due to the incurred embodied footprint) followed by a hyperbolic decrease towards a horizontal asymptote (imposed by the total operational footprint rate). The red dots indicate evaluation points at a fixed time period since the release of each new generation (six months in this paper, without loss of generality), which enables comparing how the instantaneous total carbon of the hardware family evolves over time.

III. Experimental Setup

The framework and metrics to characterize the intergenerational carbon footprint of AI infrastructure as described in the previous section can be applied to any family of AI hardware. In this paper we consider Google's versatile Tensor Processing Units (TPUs) for which Schneider et al. [9] provide data for three generations of TPUs. Table I lists and characterizes the three generations of versatile TPUs (v4i, v5e, v6e) with their release data (provided by EpochAI [3]). The embodied carbon footprint is the total footprint for manufacturing a server with eight TPUs. The operational carbon rate is the carbon emissions per month of operation computed assuming a location-based accounting method (i.e., assuming carbon intensity CI = 366 gCO_2/kWh of the local grid).[2] Normalized performance for each generation (for real Google workloads, not peak performance) is computed by dividing the total carbon emissions by the reported Compute Carbon Intensity (CCI) (gCO_2 per ExaFLOP or 10^{18} floating-point operations). CCI is a measure for the carbon efficiency of the hardware, according to which v5e and v6e are 1.17× and 2.93× more carbon-efficient than v4i, respectively.

We further recognize that the AI workloads also evolve over time, i.e., AI models become more complex which reflects itself in both training and inference. EpochAI [3] reports that between 2020 and 2025, the training compute workload (measured in ZettaFLOP or 20^{21} FLOPs) of frontier AI models increased with an average factor of 4.5× per year, with a 90% confidence interval of 3.9× to 5.2×. Unfortunately, there is less data available for inference, although Google reports a 50× increase in the number of processed tokens between May 2024 and May 2025 [7]. To account for the growth in AI model complexity, we hence leverage the available data for training (which seems more accurate and conservative compared to inference) to estimate how AI model complexity has grown over the years. Table I at the bottom reports how normalized AI complexity (i.e., number of FLOPs) has evolved alongside the three TPU generations.

IV. Results

We now explore how the intergenerational carbon footprint of AI infrastructure evolves over time while considering fixed-work and variable-work scenarios.

Fixed Work. The fixed-work scenario recognizes that successive generations of hardware are more powerful implying that one needs fewer hardware accelerators to get the same amount of work done in the same amount of time. The decrease in hardware instances is proportional to the relative increase in performance, which in turn translates into a proportional decrease in embodied and operational footprint. More specifically, this means that the embodied footprint equals

$$C_{emb,i} = \frac{EC_i}{p_i} \tag{8}$$

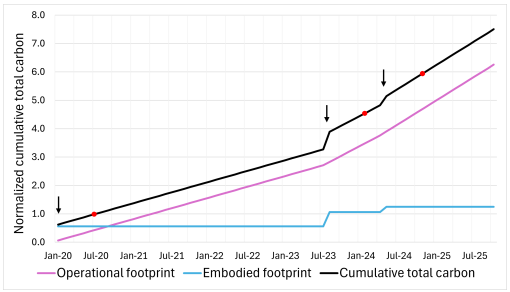
with EC_i the embodied footprint of generation i, and p_i the performance of the new generation relative to the first generation, see Table I for concrete values. Likewise, for the operational footprint rate, this translates into

$$C_{op,i}^{r} = \frac{OC_i^r}{p_i} \tag{9}$$

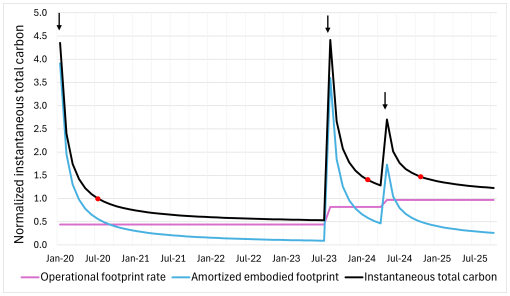
with OC_i^r the operational carbon rate for generation i, see also Table I for the various TPU generations.

Variable Work. While AI hardware becomes more powerful over the years, AI models grow more complex as well. To acknowledge this trend, we consider the variable-work scenario in which the number of hardware instances is increased to enable the new AI hardware to perform the more complex AI workload in the same amount of time. The net increase in hardware instances is proportional to the ratio of AI model growth divided by normalized AI hardware performance —

[2]We also considered Google's market-based carbon intensity CI = 135 gCO_2/kWh but this did not alter the overall conclusions.

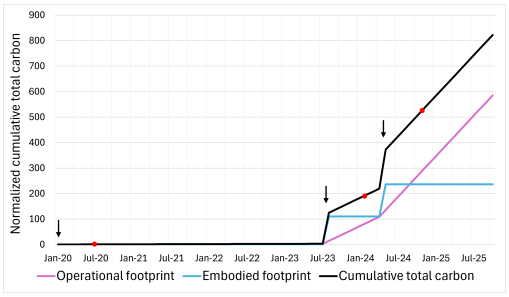


(a) Cumulative total carbon

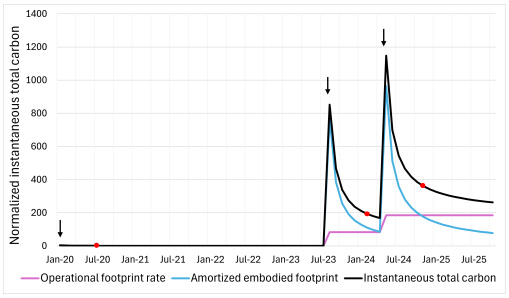


(b) Instantaneous total carbon

Fig. 2: Fixed-work scenario: cumulative and instantaneous total carbon for three Google TPU generations.



(a) Cumulative total carbon



(b) Instantaneous total carbon

Fig. 3: Variable-work scenario: cumulative and instantaneous total carbon for three Google TPU generations.

this assumes perfect workload scaling which is likely optimistic leading to an underestimation of the actual carbon footprint. For the embodied footprint, this means

$$C_{emb,i} = \frac{EC_i}{p_i} \times O_i \tag{10}$$

and likewise for the operational footprint,

$$C^r_{op,i} = \frac{OC^r_i}{p_i} \times O_i \tag{11}$$

with O_i the normalized operation count in AI model complexity, see Table I for specific values at the TPU release dates.

Analysis and Discussion. We now analyze the cumulative and instantaneous total carbon under the fixed-work and variable-work scenarios, see Figures 2 and 3, respectively. The vertical arrows denote the deployment of v4i, v5e and v6e, respectively; the red dots represent the six months after deployment; values are normalized to six months after the deployment of v4i. Under the fixed-work scenario, see Figure 2(a), we observe a 1.85× steeper slope in the cumulative total carbon upon introducing v5e because v5e has a similar operational carbon rate and a performance benefit of only 1.16× compared to v4i. The slope after introducing v6e is similar to v5e because the 1.85× increase in operational carbon rate is overcompensated by the 4.62× performance gain which in turn compensates for deploying v6e. Overall, the cumulative total carbon increases by 5.98× from v4i to v6e.

We further note that the peaks in instantaneous total carbon are similar for v4i and v5e while being substantially lower for v6e, see Figure 2(b). This is a result of the relatively high performance boost offered by v6e compared to the previous two generations, which leads to fewer TPUs needed to do the same amount of work in the same amount of time. Nevertheless, note further that the red dots indicate that despite v6e's significantly improved carbon efficiency, the instantaneous total carbon of the hardware family continues to accumulate over time. In fact, we observe a 1.44× increase from v4i to v6e.

Under the variable-work scenario we note that the cumulative total carbon footprint increases sharply with each new generation, see Figure 3(a), by 196× and 531× with the introduction of v5e and v6e, respectively. Furthermore, the instantaneous total carbon shows increasingly higher peaks with the second and third generations, see Figure 3(b). The key reason is the substantial increase in AI model growth which largely outpaces AI hardware improvements. Indeed, while AI hardware carbon efficiency improved by 1.17× and 2.93× for v5e and v6e versus v4i, respectively, the instantaneous total carbon increases by 192× and 359×.

V. Conclusion

This paper proposed a framework for analyzing the carbon footprint of AI infrastructure by quantifying both the cumulative and instantaneous total carbon across hardware generations. Using Google's TPU carbon data, we find that carbon efficiency improvements in AI hardware are significantly outpaced by AI model growth which leads to an overall increase in carbon emissions. This suggests that AI hardware needs to become carbon-efficient faster and/or AI model growth needs to slow down substantially to curb the intergenerational footprint of AI infrastructure.

Acknowledgements

This work is supported in part by the Research Foundation Flanders (FWO) grants No. G096225N and G031826N.

REFERENCES

[1] Ember (2026) – with major processing by Our World in Data. "Lifecycle carbon intensity of electricity generation – Ember" [dataset]. Ember, "Yearly Electricity Data Europe"; Ember, "Yearly Electricity Data" [original data]. Retrieved January 27, 2026 from https://archive.ourworldindata.org/20260127-111953/grapher/carbon-intensity-electricity.html (archived on January 27, 2026).

[2] International Energy Agency (IEA): Energy and AI, 2025. URL https://www.iea.org/reports/energy-and-ai.

[3] Epoch AI. Data on ai models, 07 2025. URL https://epoch.ai/data/ai-models. Accessed: 2026-01-05.

[4] 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*, 2(9), 2021.

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

[6] OpenAI. OpenAI and NVIDIA announce strategic partnership to deploy 10 gigawatts of NVIDIA systems, 09 2025. URL https://openai.com/index/openai-nvidia-systems-partnership/. Accessed: 2026-01-29.

[7] S. Pichai. Google i/o 2025: From research to reality, 2025. URL https://blog.google/innovation-and-ai/technology/ai/io-2025-keynote/#google-beam. Accessed: 2026-01-14.

[8] A. Plotnik, K. Ganesan, N. E. Jerger, and M. C. Jeffrey. Intergenerational embodied carbon. In *Workshop on Hot Topics in Ethical Computer Systems (HotEthics)*, 2024.

[9] I. Schneider, H. Xu, S. Benecke, D. Patterson, K. Huang, P. Ranganathan, and C. Elsworth. An introduction to life-cycle emissions of artificial intelligence hardware. *IEEE Micro*, 45(5):9–19, 2025. doi: 10.1109/MM.2025.3592568.

[10] A. Shehabi, S. J. Smith, A. Hubbard, A. Newkirk, N. Lei, M. A. Siddik, B. Holecek, J. G. Koomey, E. R. Masanet, and D. A. Sartor. 2024 United States data center energy usage report – US Department of Energy – Lawrence Berkeley National Laboratory, Dec. 2024. URL https://escholarship.org/uc/item/32d6m0d1.

[11] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. A. Behram, J. Huang, C. Bai, M. Gschwind, A. Gupta, M. Ott, A. Melnikov, S. Candido, D. Brooks, G. Chauhan, B. C. Lee, H.-H. S. Lee, B. Akyildiz, M. Balandat, J. Spisak, R. Jain, M. Rabbat, and K. M. Hazelwood. Sustainable AI: Environmental implications, challenges and opportunities. In *Proceedings of the Fifth Conference on Machine Learning and Systems (MLSys)*, Aug. 2022.

[12] C.-J. Wu, B. Acun, R. Raghavendra, and K. M. Hazelwood. Beyond efficiency: Scaling AI sustainably. *IEEE Micro*, 44(5):37–46, 2024.