Re-Evaluating Storage Carbon Emissions In Machine Learning Workloads

DOROTA KOPCZYK, University of Minnesota, USA
ABHISHEK CHANDRA, University of Minnesota, USA

As machine learning (ML) workloads grow in scale, the carbon impact of data storage is underexplored. Despite the dominance of solid-state drives (SSDs) in ML pipelines for their performance benefits, the environmental trade-offs with traditional hard disk drives (HDDs) are not well understood. We compare the performance and total carbon cost of SSDs and HDDs in ML training workloads. To evaluate carbon impact driven by storage, we use the MLPerf Storage benchmark along with carbon emissions data from two energy grids. We find that although SSDs have significantly higher embodied emissions, their lower operational carbon and faster runtimes make them more efficient for I/O-bound ML workloads—especially once data exceeds memory capacity. While SSDs generally amortize their carbon cost over time, often outperforming HDDs in total emissions, there are caveats. When considering regional energy mix, results suggest that carbon-aware ML infrastructure should consider workload size, memory constraints, and grid intensity—not just device specifications.

CCS Concepts: • **Computing methodologies** → *Machine learning*; *Machine learning*; • **Hardware** → **Impact on the environment**; **Impact on the environment**; *Communication hardware, interfaces and storage.*

Additional Key Words and Phrases: Sustainable Computing, ML System, Storage Carbon Emissions, SSD vs HDD, Embodied Carbon, Operational Carbon, Carbon-aware ML Infrastructure

1 INTRODUCTION

The increasing scale and complexity of ML workloads have intensified demand for optimized computation and storage systems. While recent research and industry practices have emphasized accelerators such as GPUs to improve ML performance [10], comparatively little attention has been given to the role of storage systems in shaping carbon impact. Storage plays an important role in ML pipelines given the large quantities of data used for training and inference. Solid state drives (SSDs), with their low latency and high throughput, are now the default choice in most ML pipelines. In Google's guide "Design storage for AI and ML workloads in Google Cloud", a hard disk drive (HDD) is not mentioned [4]. Yet, this shift toward SSD-centric infrastructure may obscure environmental trade-offs that merit reevaluation.

Recent work highlights that storage contributes significantly to both embodied and operational carbon emissions [9, 12, 13, 24]. SSDs often carry up to 8x the embodied carbon of HDDs [24], but they consistently outperform HDDs in random I/O workloads—such as those observed in ML training—according to benchmarks like MLPerf Storage [15]. Whether these performance gains justify their environmental cost, especially given HDDs' lower footprint and recyclability, leads us to a number of questions on the sustainability of storage in ML workloads.

- Are SSDs a sustainable choice for ML training? How do they compare to slower lower-embodied-cost HDDs?
- How does regional grid carbon intensity affect this trade-off between SSDs and HDDs?
- Can strategies like parallelism help HDDs close the performance and/or sustainability gap with SSDs?
- How do memory constraints and GPU usage influence the carbon impact of storage on ML workloads?

We focus on the ML training phase due to its high storage and compute demands compared to inference [23, 30]. ML training workloads demand high-throughput and random-access I/O which challenges HDDs [2, 11, 16, 20]. By leveraging the MLPerf Storage benchmark [15] to simulate data access patterns along with carbon emissions data from two energy grids, we assess how storage performance and emissions interact over time. We extrapolate our results to study scenarios including parallel disks, GPUs, and memory constraints.

The key takeaways from our study are:

- SSDs surpass HDDs in total carbon efficiency over their lifetime once the training dataset exceeds memory capacity, despite a 10× higher embodied carbon cost.
- The break-even point (in terms of the time it takes for SSDs to surpass HDDs in carbon efficiency) depends on workload size, duration, and grid carbon intensity; SSDs amortize their cost faster in I/O-bound scenarios.
- Regional energy mix dramatically shifts storage carbon trade-offs: HDDs are cleaner in renewables-heavy grids, but lose ground in fossil-fuel-heavy ones.
- Parallelization helps HDDs catch up in performance but accelerates operational and embodied carbon emissions, exacerbating their long-term sustainability gap from SSDs.
- Slower storage increases GPU idle time, compounding emissions even when storage power draw is low.

We position SSDs not as universally superior to HDDs, but as conditionally preferable when their I/O performance offsets embodied emissions over time. We believe these results are an initial step to encourage further research to refine and reassess ML-storage sustainability across diverse workloads and systems.

2 BACKGROUND

Storage lifespan and utilization vary widely. In data centers, devices typically last 4–7 years for HDDs, 5–10 years for SSDs, depending on environmental and workload conditions [6]. However, device utilization data is often proprietary and under-reported [22], complicating life-cycle accounting. Studies of ML infrastructure highlight that storage can be a significant carbon sink, especially in GPU-heavy systems where CPU and storage contributions are often overlooked. Life Cycle Assessments (LCAs) help quantify both embodied and operational emissions of storage technologies, offering a more complete picture of their environmental impact [21, 24, 28, 29].

Authors' addresses: Dorota Kopczyk, University of Minnesota, Minneapolis, Minnesota, USA, kopcz004@umn.edu; Abhishek Chandra, University of Minnesota, Minneapolis, Minnesota, USA, chandra@umn.edu.

However, some manufacturers still omit carbon footprint details, limiting transparency.

Operational Carbon. Operational carbon is emissions from energy consumed during use. SSDs consume less energy during active operations compared to HDDs in ML workloads due to the lack of moving parts. These energy usage statistics are documented in manufacturer datasheets, which provide power ratings for devices.

Embodied Carbon. Embodied carbon includes emissions from manufacturing, transport, and disposal. SSDs emit ten times the embodied carbon of HDDs due to energy-intensive NAND flash production. This is due to fossil fuels in regions where semiconductor factories are concentrated. For example, SSDs produce 0.109 kg CO_2e per GB, versus 0.01 kg CO_2e per GB for an HDD [13].

Storage Benchmarks. The MLPerf Storage Benchmark, introduced by MLCommons, simulates accelerator "think times" to evaluate how well storage systems handle high-throughput, low-latency ML workloads—without requiring actual GPUs [8].

3 METHODOLOGY

We evaluate storage carbon trade-offs in ML training workloads using the MLPerf Storage Benchmark [14] from MLCommons with a popular model, ResNet-50. Introduced in 2015, ResNet-50 is a widely used deep neural network [7]. The MLPerf Storage benchmark specifically uses a tool named DLIO (Deep Learning I/O) to generate synthetic datasets in ResNet form [5]. Tests were run iteratively on a single desktop machine under identical conditions, using both HDD and SSD configurations. We ran all experiments on a single Ubuntu 22.04 machine with an Intel i5-4590 CPU (4 cores), 31 GB RAM, a SATA SSD, and a SATA HDD. An idle NVIDIA RTX 3060 was present but not used. All tests were conducted using Python 3.13 on Linux kernel 6.5.0.

3.1 Benchmark Execution and Iterative Testing

We cleared system caches between runs for consistency. We used datasets ranging from 21GB to 53GB, with memory fixed at 31GB to capture both memory and I/O-bound scenarios.

To estimate power use, we recorded active and idle durations and combined them with device states (idle, read, write). Table 1 summarizes the device specifications.

Table 1. Power Consumption of Storage Devices

Storage Device	Type	Capacity	Idle Watts	Read Watts	Write Watts
LITEONIT LCS-256M6S	SSD	256GB	3.5	7.0	4.0
Seagate ST1000DM003-1ER162	HDD	1TB	3.36	5.9	5.9

These power ratings, combined with observed idle and active durations, were used to estimate energy consumption and compute per-run operational carbon emissions.

3.2 Active vs Idle Storage State

We estimated energy consumption by measuring the fraction of time each storage device spent active versus idle during benchmark runs. Standard tools like `iostat` sample too coarsely, often missing short I/O bursts. We sampled `/proc/diskstats` at 4 ms intervals which matched the kernel tick rate. We tracked changes in read/write counts to classify device activity.

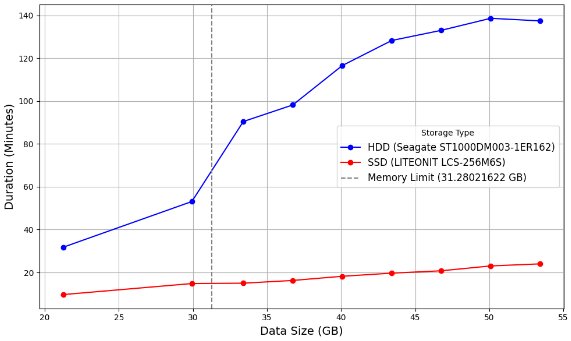

Fig. 1. Duration of Simulated Training Runs - HDD vs SSD

These activity proportions, combined with manufacturer-reported power values, enabled per-run estimates of storage energy use. We use those to extrapolate carbon emissions.

3.3 Carbon Emissions

We calculate operational carbon by estimating total energy consumed during each benchmark run and scaling it by the local grid's carbon intensity, as shown in Equation (1):

$$\text{Emissions} = \text{Energy Consumed} \times \text{Carbon Intensity} \qquad (1)$$

Energy is computed from idle, read, and write durations, using power ratings from device datasheets. This isolates emissions attributable to storage during ML workloads.

We assume a baseline carbon intensity of 0.364 kg CO_2e/kWh for a fossil-heavy Midwestern U.S. state, Minnesota [26]. In contrast, a solar-rich Western U.S. state averages 0.163 kg CO_2e/kWh, California [27]. Clean grid periods can reach intensities as low as 0.011 kg CO_2e/kWh for wind and 0.012 for nuclear [1]. While we use average carbon intensity values for each region, actual emissions can vary significantly with time of day and season due to changes in renewable availability and load demand [25]. We discuss the impact of such variations later.

4 RESULTS

We focus on a few outcomes: the runtime differences between SSDs vs HDDs and how those differences translate into operational and total carbon emissions.

4.1 Duration SSD vs HDD

Figure 1 shows the average duration of simulated training runs using HDD and SSD across dataset sizes from 21GB to 53GB. A divergence appears at the 31.28GB memory threshold (dashed line). SSD durations increase modestly from 12 to 25 minutes as data grows, reflecting stable low-latency I/O. In contrast, HDD durations jump sharply—from 33 to 138 minutes—once data exceeds memory, showing a shift from memory-bound to I/O-bound behavior. This inflection is driven by HDD seek delays and rotational latency, which compound as I/O demands grow. This gap is especially pronounced

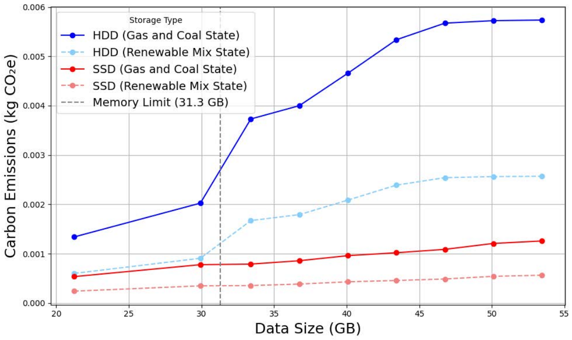

Fig. 2. Storage Device Operational Carbon

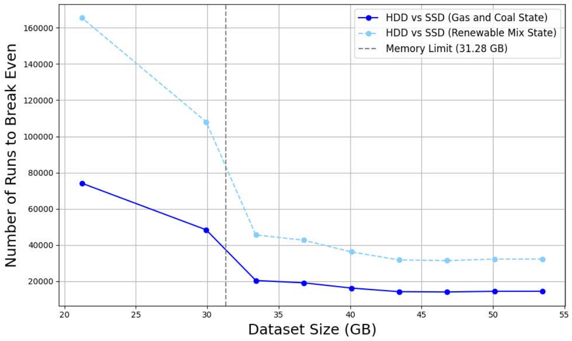

Fig. 3. Break-Even Runs by Dataset Size and SSD Runtime

in early epochs, when OS-level caching has not yet taken effect. Even though the full dataset may eventually reside in DRAM, the benchmark issues randomized reads that incur stalls during the initial epoch, leading to low GPU utilization until memory is fully populated.

4.2 Carbon Emissions

As shown in Figure 2, once data size exceeds the memory limit, HDD operational emissions increase sharply while SSD emissions remain relatively stable. This is due to the HDD's longer runtimes under I/O-bound workloads, where modest power efficiency cannot offset prolonged activity.

Although SSDs emit more embodied carbon, operational efficiency can shift the total emissions balance over time. To estimate embodied impact, we use the Storage Embodied Factor (SEF), which quantifies manufacturing, transport, and end-of-life emissions per terabyte [24]. In our analysis we use SEF values of SSDs producing 0.109 kg CO_2e per GB, compared to 0.01 kg CO_2e per GB for HDDs in their manufacture [13]. SSDs exhibit significantly higher SEF values due to the carbon-intensive NAND flash fabrication process. Much of this is derived from fossil fuels in regions where semiconductor factories are concentrated.

This gap means SSDs start with a higher carbon "debt," but their performance can amortize that cost. The break-even point, where cumulative emissions from HDDs surpass those from SSDs, depends on workload.

We selected the 53GB workload because it exceeds memory capacity, making the job I/O-bound and reflective of real ML training scenarios. In this setting, SSDs—despite an order-of-magnitude higher embodied carbon—eventually surpass HDDs in total carbon efficiency. The break-even point occurs after 14,440 runs in a fossil-heavy grid and 32,223 runs in a renewables-heavy grid. This crossover reflects the compounding effect of HDDs' slower runtimes and higher per-run emissions. Notably, the slope of cumulative emissions for HDDs is 3–4x steeper than that of SSDs, due to prolonged active durations under load. In practice, for the 53GB workload, this means SSDs "catch up" after roughly eight months of continuous training for the fossil-heavy grids, and around 17 months for the clean-energy grid. Both of these durations are much less compared to the typical lifetime of an SSD (5-10 years). In contrast, for a 21GB workload that fits in memory, the break-even point would occur after about 74,000 runs in a fossil-heavy grid which, thanks to an SDD duration a third of the HDD's, would take 16.5 months in a fossil-heavy grid.

We can see the number of runs to reach a similar break-even point across all data sizes in Figure 3. This shows how break-even runs shrink as data size increases beyond memory. The exact break-even point shifts with grid intensity: cleaner electricity delays parity, while dirtier grids accelerate SSD carbon payback.

5 BEYOND SINGLE DISKS

While our core analysis compares individual SSD and HDD devices, we also explore how parallel disk configurations and accelerator interactions affect overall emissions.

5.1 Considering Parallel HDDs

While HDDs exhibit lower operational carbon emissions and a significantly smaller embodied carbon footprint compared to SSDs, their performance limitations—particularly under I/O-bound workloads like our 53GB dataset—create a significant bottleneck. One remedy is parallelization which stripes data across multiple drives to enable concurrent reads and improve throughput [19].

In ideal conditions, a parallel disk like RAID0 scales throughput nearly linearly with the number of disks. Because our benchmark workload is read-only, this setup provides a best-case scenario for HDD performance gains. To estimate how many drives would be required to match SSD throughput, we measured training I/O throughput across different parallel HDD configurations and identified the smallest number of HDDs that approximated the SSD's baseline performance. For the 53GB workload, that threshold was reached with four parallel HDDs.

While this configuration reduced per-run duration and improved I/O efficiency, it also increased *both* embodied and operational carbon due to the added hardware. In other words, performance scaled, but so did emissions, making parallel HDD setups less favorable over time compared to high-performing SSDs, especially under frequent, I/O-heavy workloads.

Comparing the total carbon emissions from SSDs and a RAID0 configuration of four HDDs (HDDx4) over repeated runs of a 53GB workload, we find that in the fossil fuel-heavy state, RAID0 HDDs

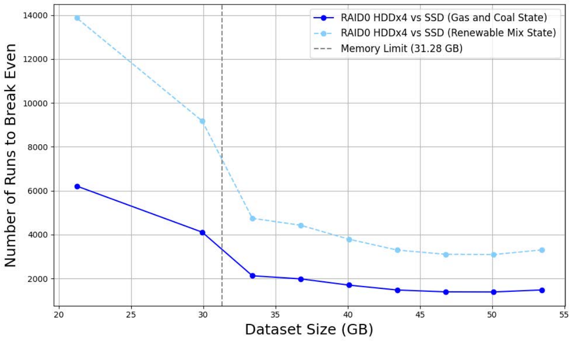

Fig. 4. Break-Even Runs vs Dataset Size (RAID0 HDDx4)

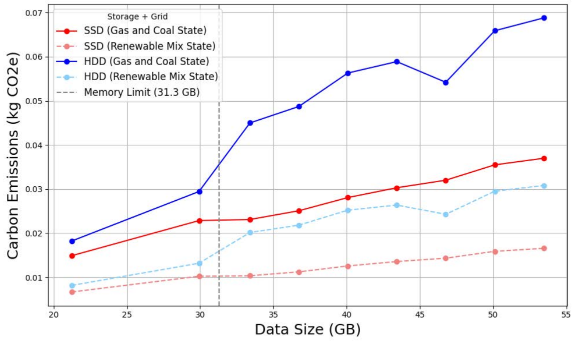

Fig. 5. Operational Carbon With A100 GPU

surpass SSDs in total emissions after just 1,478 runs, at which point both configurations have emitted approximately 71.8 kg CO_2e. However, in the cleaner renewable mix state, that break-even point shifts substantially: SSDs catch up after 3,075 runs, despite their higher embodied carbon, compared to their prior 32,223 runs in the single-disk scenario.

To extend this analysis, we compare break-even points across datasets of varying sizes to show how storage performance and operational emissions shape carbon trade-offs over time. For example, at just 21GB, SSDs amortize their embodied carbon much more efficiently, requiring over 6,200 runs to break even with RAID0 HDDs, compared to over ten times as much in the single-HDD scenario. When switching to a cleaner grid mix, the entire break-even curve shifts. In the renewables-heavy state, break-even points stretch dramatically, about 14,000 runs for the smallest dataset size. This again highlights the interplay between workload size, system configuration, and grid intensity: cleaner grids delay parity, while dirtier grids accelerate SSD carbon payback—even in RAID configurations. (Figure 4).

5.2 Storage and GPU Embodied Carbon

Although our primary focus is storage, we briefly examine how it interacts with GPUs to influence operational carbon. Slow storage can cause GPU idle time, reducing overall efficiency and increasing emissions—even in single-node deployments. For our study, we consider the NVIDIA A100 GPU, which draws approximately 250W under PCIe configurations [17]. Since idle power is not reliably documented, we conservatively assume 25W idle and 250W active for modeling. This enables a simplified but effective estimate of GPU emissions during storage-bound phases.

To estimate GPU-related operational carbon, we use MLPerf's Training Accelerator Utilization metric, which captures the fraction of time the simulated accelerator is active. We model the A100 GPU drawing 250W during active phases and 25W while idle, accounting for stalls caused by slower storage. We exclude GPU embodied carbon, as it is constant across both SSD and HDD scenarios and does not affect relative comparisons.

We integrate these emissions with those from storage, adding on the following:

$$\text{Emissions}_{\text{GPU}} = \left(\frac{P_{\text{active}} \cdot t \cdot U + P_{\text{idle}} \cdot t \cdot (1 - U)}{3600} \right) \cdot \text{CI}$$

- P_{active} - Active power draw of the GPU (in watts)
- P_{idle} - Idle power draw of the GPU (in watts)
- t - Duration of the workload (in seconds)
- U - Utilization fraction of the GPU (from MLPerf output)
- CI - Carbon intensity of electricity (in kg CO_2e per kWh)

Carbon intensity is based on either the fossil-heavy state's or the clean state's grid average (0.364 kg CO_2e/kWh, 0.163 kg CO_2e/kWh).

Figure 5 shows total operational carbon emissions per training run from combined storage and GPU usage, broken down by dataset size and grid carbon intensity. At smaller dataset sizes (e.g., 21GB), emissions are comparable across devices and grids, ranging from 0.002 to 0.005 kg CO_2e. However, beyond the 31GB memory threshold, the emissions gap widens sharply. For the 53GB dataset under a fossil-heavy grid, HDD+GPU emissions reach 0.124 kg CO_2e per run, while SSD+GPU emissions stay at 0.039 kg CO_2e—a more than 3x difference. In the renewables-heavy grid, absolute emissions are lower across the board, but the pattern holds: HDD+GPU emits 0.056 kg CO_2e, versus 0.018 kg CO_2e for SSD+GPU at 53GB, again a 3x gap.

6 DISCUSSION

Our results reveal that storage carbon trade-offs in ML training workloads are highly dependent on workload size, energy mix, and device configuration. We next highlight implications for deployment strategies, memory constraints, and system interactions.

6.1 Storage Tradeoffs by Grid Region

While embodied carbon is a fixed upfront cost, operational carbon in datacenters varies significantly with grid energy sources. For example, the renewable-heavy state's grid is nearly twice as clean as the fossil-heavy state's, making identical ML training jobs substantially less carbon-intensive when executed in cleaner regions.

Table 2 shows the average operational carbon emissions per training run with GPU across a range of data sizes. This highlights how much carbon efficiency depends on the underlying energy mix. At

Table 2. Operational Carbon Emissions per Run (kg CO_2e)

Energy Source	**21.24GB**		**53.44GB**	
	SSD	HDD	SSD	HDD
Wind	0.0005	0.0006	0.0011	0.0021
Nuclear	0.0005	0.0006	0.0012	0.0023
Solar	0.0017	0.0021	0.0042	0.0077
Renewables State Avg	0.0067	0.0082	0.0166	0.0308
Gas+Coal State Avg	0.0149	0.0182	0.0370	0.0688
Natural Gas	0.0201	0.0245	0.0498	0.0926
Oil	0.0266	0.0325	0.0660	0.1228
Coal	0.0336	0.0411	0.0833	0.1550

21.24 GB, for instance, the emissions range from 0.0005 kg CO_2e (wind-powered SSD) to over 0.0336 kg CO_2e (coal-powered SSD), a nearly 67x difference for the same hardware and workload.

These differences scale with larger data sizes and persist across both SSD and HDD configurations. No hardware choice is carbon-efficient in a vacuum; its sustainability depends heavily on where and when the job is executed.

This variability suggests an important design opportunity: carbon-aware workload shifting [1]. ML training jobs could be scheduled during lower-carbon times of day or routed to datacenters in cleaner regions to reduce emissions without any hardware change. In this view, temporal and geographic flexibility may substitute for more carbon-intensive hardware upgrades, especially in storage-bound workloads where emissions vary by both device and grid context.

6.2 Memory as a Pivot Point in Trade-offs

An inflection point emerges in our results at the memory boundary (31GB), where workloads shift from being memory-bound to I/O-bound. Below this threshold, datasets can be cached in RAM, minimizing disk activity and leveling the carbon impact of storage types. However, once data exceeds memory, disk performance becomes critical—especially for random-access ML training workloads. At this point, SSDs rapidly amortize their higher embodied carbon due to shorter runtimes, higher throughput, and lower GPU idle time.

Systems with limited RAM may benefit from more aggressive use of multi-tiered storage hierarchies, for example RAM for hot data, SSDs for warm data, and HDDs for cold storage. However, when active datasets cannot be held in memory, the carbon efficiency of SSDs becomes a clear advantage. Future ML infrastructure may need to balance memory, caching policy, and storage device choice holistically to minimize emissions across workloads of varying size.

Systems with limited RAM may benefit from more aggressive use of multi-tiered storage hierarchies—for example, RAM for hot data, SSDs for warm data, and HDDs for cold storage. Intermediate caches, such as application-level prefetch layers, could further smooth these trade-offs by absorbing bursty access patterns seen when loading training data. However, when active datasets cannot be held in memory, the carbon efficiency of SSDs becomes a clear advantage. Future ML infrastructure may need to balance memory, caching policy, and storage device choice holistically to minimize emissions across workloads of varying size.

6.3 Parallelism Under Carbon Constraints

Parallelism offers a theoretical way to close the performance gap between HDDs and SSDs by distributing across multiple disks. This improves HDD throughput and reduces runtime, but scales up both embodied and operational emissions. Each added disk increases energy use and hardware footprint, causing the carbon break-even point with SSDs to arrive much sooner than in the single-disk case. Parallelism boosts performance while accelerating the emissions penalty.

These results suggest system configurations should align with workload and local carbon intensity. In clean grids or for infrequent, latency-tolerant training, parallel HDDs may still offer some performance with lower upfront cost. In fossil-heavy grids or under frequent pressure, SSDs amortize their higher embodied emissions.

6.4 Limitations and Future Work

While our results reveal meaningful trends in storage-related emissions, they are based on a single-machine setup using a synthetic benchmark workload and specific hardware. Moreover, we relied on device datasheets to estimate energy usage rather than direct power measurements. As such, we caution against overgeneralization without further empirical validation. Still, this analysis provides a useful lens on underexplored trade-offs between storage performance and sustainability, and underscores the need for more carbon-aware benchmarking tools.

Future work should extend beyond SSD and HDD comparisons to include full-system interactions—encompassing memory configurations, accelerator power profiles, and data transfer bottlenecks. For instance, prolonged I/O durations can extend GPU runtime, inflating emissions even with low-power storage. To avoid idle stalls and fully utilize accelerators, storage must deliver data at sufficient throughput. This is especially relevant as newer GPUs such as the NVIDIA H100 can draw up to 500W [18], compounding the emissions impact of storage delays.

Broader analyses should benchmark carbon performance across larger-scale systems, including inference workloads and models beyond ResNet-50, such as large language models (LLMs). Ultimately, future studies should seek configurations that balance performance and sustainability across real-world ML deployments.

7 RELATED WORK

Prior work has highlighted storage as a nontrivial contributor to carbon emissions in data centers [9, 12, 13, 24]. Tannu and Nair [24] show that SSDs have significantly higher embodied carbon than HDDs due to energy-intensive fabrication. McAllister et al. [13] call for better modeling of storage emissions and carbon-aware design tools.

Recent ML-focused studies examine the carbon impact of GPU use and training location [1, 30], but few isolate the role of storage in shaping emissions. HPCwire [8] notes that storage bottlenecks are increasingly a limiting factor in AI. While some work explores I/O bottlenecks in ML pipelines [2, 16], our study takes a second look at classic storage devices to directly benchmark SSD and HDD carbon trade-offs under ML-specific storage patterns [15]. With recent work demonstrating that coordinated downthrottling of GPUs can reduce power draw during underutilized periods [3], it would be compelling to extend this approach to I/O bottlenecks, where slow storage similarly leaves accelerators idle.

8 CONCLUSION

This paper examines the carbon trade-offs between SSDs and HDDs in ML training workloads, highlighting how performance and energy use interact with embodied emissions. While SSDs have significantly higher embodied carbon due to energy-intensive manufacturing, their superior I/O performance enables shorter runtimes, lower operational emissions, and less GPU idle time—especially when workloads exceed memory capacity.

In contrast, HDDs offer lower embodied carbon and cost but struggle under I/O-bound workloads, leading to longer runtimes and higher cumulative emissions over time. Even with parallelization, the additional hardware increases both embodied and operational carbon, often accelerating the break-even point where SSDs become cleaner overall.

Our results suggest that SSDs are not universally better, but they are conditionally preferable in many ML training scenarios. The optimal storage choice depends on workload size, memory constraints, usage frequency, and local energy mix. Ultimately, carbon-aware infrastructure decisions must consider not just device specifications, but system-wide dynamics over time.

This single-machine study provides a starting point for more nuanced evaluation of storage in sustainable ML systems. Future work should explore these trade-offs at scale, incorporating distributed storage stacks, varied GPU configurations, and real-world workloads.

ACKNOWLEDGMENTS

This work was supported in part by Seagate Technology and we thank them for their support.

REFERENCES

[1] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '23)*. ACM, 118–132. https://doi.org/10.1145/3575693.3575754

[2] Alex Aizman, Gavin Maltby, and Thomas Breuel. 2020. High Performance I/O For Large Scale Deep Learning. arXiv:2001.01858 [cs.DC] https://arxiv.org/abs/2001.01858

[3] Jae-Won Chung, Yile Gu, Insu Jang, Luoxi Meng, Nikhil Bansal, and Mosharaf Chowdhury. 2024. Reducing Energy Bloat in Large Model Training. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles (SOSP '24)*. ACM, 144–159. https://doi.org/10.1145/3694715.3695970

[4] Google Cloud. 2023. Design storage for AI and ML workloads in Google Cloud. https://cloud.google.com/architecture/ai-ml/storage-for-ai-ml

[5] Hariharan Devarajan, Huihuo Zheng, Anthony Kougkas, Xian-He Sun, and Venkatram Vishwanath. 2021. DLIO: A Data-Centric Benchmark for Scientific Deep Learning Applications. In *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. 81–91. https://doi.org/10.1109/CCGrid51090.2021.00018

[6] Don Hall. 2023. Life expectancy of a drive: HDD, SSD, and flash. https://www.enterprisestorageforum.com/hardware/life-expectancy-of-a-drive/

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] https://arxiv.org/abs/1512.03385

[8] HPCwire. 2024. GenAI: It's Not the GPUs, It's the Storage. *HPCwire*. https://www.hpcwire.com/2024/09/12/genai-its-not-the-gpus-its-the-storage/ Accessed: 2024-12-17.

[9] Shixin Ji, Zhuoping Yang, Xingzhen Chen, Stephen Cahoon, Jingtong Hu, Yiyu Shi, Alex K. Jones, and Peipei Zhou. 2024. SCARIF: Towards Carbon Modeling of Cloud Servers with Accelerators. arXiv:2401.06270 [cs.DC] https://arxiv.org/abs/2401.06270

[10] Sameer Kumar, James Bradbury, Cliff Young, Yu Emma Wang, Anselm Levskaya, Blake Hechtman, Dehao Chen, HyoukJoong Lee, Mehmet Deveci, Naveen Kumar, Pankaj Kanwar, Shibo Wang, Skye Wanderman-Milne, Steve Lacy, Tao Wang, Tayo Oguntebi, Yazhou Zu, Yuanzhong Xu, and Andy Swing. 2021. Exploring the limits of Concurrency in ML Training on Google TPUs. arXiv:2011.03641 [cs.LG] https://arxiv.org/abs/2011.03641

[11] Xiaqing Li, Guangyan Zhang, H. Howie Huang, Zhufan Wang, and Weimin Zheng. 2016. Performance Analysis of GPU-Based Convolutional Neural Networks. In *2016 45th International Conference on Parallel Processing (ICPP)*. 67–76. https://doi.org/10.1109/ICPP.2016.15

[12] Yueying (Lisa) Li, Omer Graif, and Udit Gupta. 2024. Towards Carbon-efficient LLM Life Cycle. https://hotcarbon.org/assets/2024/pdf/hotcarbon24-final154.pdf

[13] Sara Mcallister, Fiodar Kazhamiaka, Daniel S. Berger, Rodrigo Fonseca, Kali Frost, Aaron Ogus, Maneesh Sah, Ricardo Bianchini, George Amvrosiadis, Nathan Beckmann, and Gregory R. Ganger. 2025. A Call for Research on Storage Emissions. *SIGENERGY Energy Inform. Rev.* 4, 5 (April 2025), 67–75. https://doi.org/10.1145/3727200.3727211

[14] ML Commons. 2024. New MLPerf Storage v1.0 Benchmark Results Show Storage Systems Play a Critical Role in AI Model Training Performance. https://mlcommons.org/2024/09/mlperf-storage-v1-0-benchmark-results/. Accessed: 2025-06-26.

[15] MLCommons. 2023. MLPerf Storage Benchmark. (2023). https://github.com/mlcommons/storage

[16] NVIDIA. 2024. Machine Learning Frameworks Interoperability, Part 2: Data Loading and Data Transfer Bottlenecks. *NVIDIA Technical Blog*. https://developer.nvidia.com/blog/machine-learning-frameworks-interoperability-part-2-data-loading-and-data-transfer-bottlenecks/ Accessed: 2024-10-17.

[17] NVIDIA Corporation. 2023. NVIDIA A100 Tensor Core GPU Datasheet. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet.pdf Accessed: 2025-05-05.

[18] Emmanuel Ohiri. 2023. Nvidia A100 versus H100: How do they compare? https://www.cudocompute.com/blog/comparative-analysis-of-nvidia-a100-vs-h100-gpus#power-efficiency-and-environmental-impact

[19] D.A. Patterson, P. Chen, G. Gibson, and R.H. Katz. 1989. Introduction to redundant arrays of inexpensive disks (RAID). In *Digest of Papers. COMPCON Spring 89. Thirty-Fourth IEEE Computer Society International Conference: Intellectual Leverage*. 112–117. https://doi.org/10.1109/CMPCON.1989.301912

[20] Sebastian Ruder. 2017. An overview of gradient descent optimization algorithms. arXiv:1609.04747 [cs.LG] https://arxiv.org/abs/1609.04747

[21] Seagate Technology LLC. 2020. Understanding Life Cycle Assessment and Embodied Carbon. https://www.seagate.com/resources/understanding-life-cycle-assessment-and-embodied-carbon/. Accessed: 2025-05-15.

[22] Arman Shehabi, Sarah J. Smith, Alex Hubbard, Alex Newkirk, Nuoa Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024. *2024 United States Data Center Energy Usage Report*. Technical Report. Lawrence Berkeley National Laboratory. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report_1.pdf Accessed: 2024-12-17.

[23] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1355

[24] Swamit Tannu and Prashant J. Nair. 2023. The Dirty Secret of SSDs: Embodied Carbon. *ACM SIGEnergy Energy Informatics Review* 3, 3 (Oct. 2023), 4–9. https://doi.org/10.1145/3630614.3630616

[25] John Thiede, Noman Bashir, David Irwin, and Prashant Shenoy. 2023. Carbon Containers: A System-level Facility for Managing Application-level Carbon Emissions. In *Proceedings of the 2023 ACM Symposium on Cloud Computing* (Santa Cruz, CA, USA) *(SoCC '23)*. Association for Computing Machinery, New York, NY, USA, 17–31. https://doi.org/10.1145/3620678.3624644

[26] U.S. Energy Information Administration. 2024. Minnesota State Electricity Profile. https://www.eia.gov/electricity/state/Minnesota/ Accessed: 2024-12-17.

[27] U.S. Energy Information Administration. 2025. California State Electricity Profile. https://www.eia.gov/electricity/state/California/ Accessed: 2025-05-15.

[28] Western Digital. 2023. Life Cycle Assessment: Western Digital Green SATA SSD. https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/collateral/analyst-report/life-cycle-assesment-wd-green-sata-ssd-western-digital.pdf. Accessed: 2025-05-15.

[29] Western Digital. 2023. Life Cycle Assessment: Western Digital Ultrastar DC HC670 HDD. https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/collateral/analyst-report/life-cycle-

assesment-ultrastar-dc-hc670-hdd-western-digital.pdf. Accessed: 2025-05-15.

[30] Mark Zhao, Niket Agarwal, Aarti Basant, Buğra Gedik, Satadru Pan, Mustafa Ozdal, Rakesh Komuravelli, Jerry Pan, Tianshu Bao, Haowei Lu, Sundaram Narayanan, Jack Langman, Kevin Wilfong, Harsha Rastogi, Carole-Jean Wu, Christos Kozyrakis, and Parik Pol. 2022. Understanding data storage and ingestion for large-scale deep recommendation model training: industrial product. In *Proceedings of the 49th Annual International Symposium on Computer Architecture (ISCA '22)*. ACM, 1042–1057. https://doi.org/10.1145/3470496.3533044