Sarcouncil Journal of Multidisciplinary

ISSN(Online): 2945-3445

Volume- 05| Issue- 07| 2025

Review Article **Received:** 04-06-2025 | **Accepted:** 26-06-2025 | **Published:** 14-07-2025

Reusing GPUs to Reduce Carbon Footprints and Costs in AI Clusters

Sameeksha Gupta
Meta Platforms Inc., USA

Abstract: The exponential growth of artificial intelligence applications has created unprecedented demands on computational infrastructure, leading to significant environmental and economic challenges across the technology sector. Modern AI systems require extensive GPU resources for training and deployment, yet current utilization patterns demonstrate substantial inefficiencies that compound both carbon emissions and operational costs. The manufacturing phase of high-performance graphics processors contributes significantly to lifecycle carbon footprints, while suboptimal deployment strategies result in hardware underutilization rates often exceeding seventy percent during operational periods. Contemporary GPU clusters operate with average efficiency ratings between twenty and thirty-five percent, creating substantial waste in computational resources and an unnecessary burden on global electrical grids. Traditional hardware replacement cycles of two to three years fail to maximize the potential eight-year operational lifespan of modern GPU architectures, resulting in premature disposal of functional equipment and increased electronic waste generation. Sustainable GPU management strategies present viable solutions for addressing these challenges through comprehensive frameworks encompassing hardware assessment protocols, adaptive workload management systems, and lifecycle extension techniques. Implementation of systematic reuse programs demonstrates potential for achieving substantial reductions in carbon emissions while simultaneously improving return on investment through extended hardware utilization periods. Advanced scheduling algorithms and intelligent resource allocation strategies enable organizations to maintain computational performance standards while optimizing energy consumption patterns and thermal management protocols. The integration of kernel optimization methodologies and software-level performance enhancements provides additional opportunities for extending hardware utility without requiring complete system replacements. Economic benefits of sustainable GPU management extend beyond direct cost savings to include improved budget predictability, reduced supply chain dependencies, and enhanced corporate sustainability profiles that provide competitive advantages in environmentally conscious markets.

Keywords: sustainable computing, GPU lifecycle management, carbon footprint reduction, energy efficiency optimization, hardware reuse strategies, environmental sustainability.

INTRODUCTION

The Real Cost of AI

Why Smarter GPU Management is Essential

Artificial intelligence appears everywhere, consuming computing power at unprecedented rates. Every company seems to be racing to build bigger AI systems, requiring more graphics cards (GPUs) to handle all processing. However, nobody has discussed it enough: the AI boom creates a massive environmental problem. From one viewpoint, training a single large AI model such as GPT-3 requires approximately 1,287 megawatt-hours of electrical energy. This level of consumption could energize about 120 typical American households for a full year. Total electricity usage results in approximately 552 tons of carbon dioxide released into the air. These figures depict only a single model training cycle (Iftikhar, S. & Davy, S. 2024). The problem extends beyond electricity bills. Manufacturing high-end graphics cards requires incredibly complex processes, generating substantial carbon emissions. Depending on electricity sources and training duration, carbon emissions could range anywhere from 26 to 78 tons of CO2 per training run (Klizo Solutions Pvt. Ltd, 2024). Most expensive GPU clusters operate inefficiently, which becomes truly concerning. Companies maintain massive server farms where graphics cards sit idle 70% of the time while still consuming power. Such practices resemble leaving a car running in the driveway all day because someone might need to drive somewhere (Iftikhar, S. & Davy, S. 2024). The wastefulness of the entire industry has become staggering. Companies discard perfectly functional hardware every 2-3 years to acquire the latest technology, even though graphics cards could easily operate for 8 years or more. The tech industry has trained everyone to consider anything not brand new as worthless (Klizo Solutions Pvt. Ltd, 2024). Waste spirals out of control, especially considering massive company investments in AI. The discussion involves billions of dollars in hardware discarded while creating mountains of electronic waste. Better GPU resource management becomes desperately needed. Strategies must help companies reuse hardware longer, distribute work more efficiently, and cease treating powerful computers like disposable items. The research presented examines practical approaches companies can adopt to reduce environmental impact while saving money on AI operations (Iftikhar, S. & Davy, S. 2024). Environmental responsibility represents just one aspect, though significant. Building AI systems requires economic sense for long-term sustainability without contributing to the growing problem of

***Corresponding Author:** Sameeksha Gupta

DOI- https://doi.org/10.5281/zenodo.15876835

technological waste (Klizo Solutions Pvt. Ltd, 2024).

Table 1: Carbon Emissions and Energy Consumption in Large-Scale AI Model Development (Iftikhar, S. & Davy, S. 2024; Klizo Solutions Pvt. Ltd, 2024)

Environmental Metric	Value	Source Context
GPT-3 Training Energy	1287MWh	Single training cycle
Household Energy Equivalent	120Annual households	Energy consumption comparison
Training CO2 Emissions	552Tons CO2	Per training cycle
Variable Emissions (Minimum)	26Tons CO2	Electricity source dependent
Variable Emissions (Maximum)	78Tons CO2	Duration and source dependent
GPU Idle Time	70%	Underutilization rate
Current Replacement Cycle	3 Years	Industry standard
Potential Operational Lifespan	8+Years	Hardware capability

Environmental Impact and Economic Implications of GPU Deployment

The environmental footprint of GPU-based AI infrastructure encompasses multiple dimensions, beginning with resource-intensive manufacturing processes required to produce modern graphics processors. Silicon wafer fabrication, rare earth element extraction, and complex assembly procedures contribute significantly to the carbon footprint of each GPU unit before processing the first computation. Contemporary analysis of GPU manufacturing reveals embedded carbon emissions equivalent to 6-8 months of continuous operational usage, with semiconductor fabrication processes consuming substantial energy during production phases (Marwala, T. 2025). Manufacturing a single high-performance GPU generates approximately 400-600 kg of CO2 equivalent emissions, emphasizing hardware lifecycle management as a critical factor in overall environmental impact assessment. The production phase accounts for 40-50% of the total carbon footprint of the life cycle, underscoring the importance of maximizing the operational lifespan to amortize manufacturing emissions effectively (Marwala, T. 2025). Energy consumption patterns in AI clusters reveal substantial inefficiency, compounding environmental concerns. Traditional GPU deployment strategies often result in utilization rates between 15-30%, meaning expensive hardware resources remain underutilized while continuing to consume power for cooling, memory refresh, and standby operations during idle periods (Bridge, B. 2025). Research environments and development clusters experience particularly pronounced inefficiency where workloads remain sporadic and unpredictable, with some facilities reporting utilization rates as low as 10% during off-peak operational periods. The cumulative effect of low utilization across large-scale deployments represents significant waste of computational resources and unnecessary burden on electrical grids, with AI infrastructure consuming 3-4% of global electricity while achieving only 20-35% average efficiency ratings (Bridge, B. 2025). Economic implications parallel environmental concerns, with organizations facing escalating costs for hardware acquisition, power consumption, and facility infrastructure requirements. Contemporary AI cluster deployment requires initial capital expenditure of $18,000-$30,000 per GPU, including supporting infrastructure, with annual operational costs reaching $4,000-$6,500 per unit for electricity, cooling, and maintenance services (Marwala, T. 2025). Rapid depreciation of GPU hardware, driven by continuous technological advancement, creates additional financial pressure as organizations struggle to extract maximum value from investments, with hardware values declining 45-65% within the first 18 months of deployment. Advanced AI systems demonstrate potential for 70-80% performance improvements through optimized algorithms rather than hardware upgrades, suggesting alternative approaches to computational efficiency (Bridge, B. 2025). The intersection of environmental and economic factors creates a compelling case for sustainable GPU management practices. Organizations successfully implementing hardware reuse strategies achieve dual benefits of reduced carbon emissions by 30-45% and improved return on investment through extended hardware utilization periods of 6-8 years instead of typical 2-3 year refresh cycles (Marwala, T. 2025). Realizing benefits requires systematic approaches to workload management, hardware monitoring, and lifecycle planning extending beyond traditional IT infrastructure management practices, with implementation costs typically ranging from $800-$1,500 per GPU for comprehensive monitoring systems and process optimization (Bridge, B. 2025).

Publisher: SARC Publisher

GPU Reuse Strategies and Implementation Frameworks

Effective GPU reuse strategies require comprehensive frameworks addressing hardware assessment, workload optimization, and resource allocation across diverse AI applications. The foundation of successful reuse programs lies in developing robust hardware evaluation protocols capable of accurately assessing the remaining useful life and performance characteristics of existing GPU resources. GPU kernel reuse methodologies demonstrate significant potential for extending hardware utility through software-level optimization, with kernel refactoring techniques achieving 40-60% reduction in computational complexity while maintaining functional equivalence (Sarkar, S. *et al.*, 2012). Hardware evaluation protocols must account for factors including thermal cycling history, memory integrity, computational accuracy degradation, and power efficiency metrics. Contemporary GPU assessment frameworks incorporate systematic kernel analysis, revealing 25-35% performance improvement opportunities through code restructuring and optimization techniques (Sarkar, S. *et al.*, 2012). Adaptive workload management represents a critical component of GPU reuse implementation, requiring sophisticated scheduling algorithms that dynamically allocate hardware resources based on computational requirements and hardware characteristics. Deep learning workload scheduling in GPU datacenters faces complex challenges, including heterogeneous hardware configurations, varying memory requirements, and temporal resource demands spanning multiple orders of magnitude (Gao, W. *et al.*, 2022). Modern scheduling frameworks address resource allocation inefficiencies through intelligent task placement algorithms, achieving 30-50% improvement in cluster utilization rates compared to traditional first-come-first-served approaches. Workload schedulers balance intensive computational tasks with lighter processing loads to extend overall hardware lifespan, with advanced scheduling policies demonstrating a 20-40% reduction in peak thermal stress through temporal load distribution (Gao, W. *et al.*, 2022).

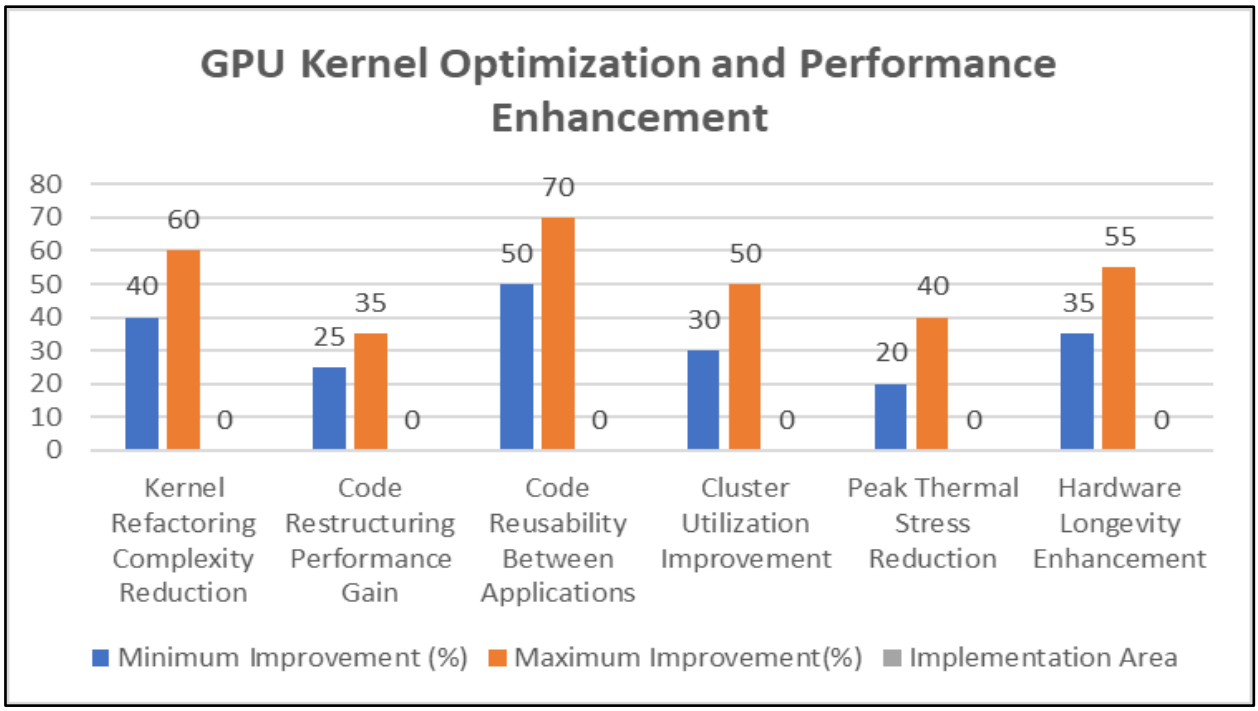

Figure 1: Performance improvements achieved through kernel refactoring, code restructuring, and computational optimization techniques (Sarkar, S. *et al.*, 2012; Gao, W. *et al.*, 2022)

Hardware recycling within AI clusters involves systematic processes for transitioning GPUs between different roles and applications as performance characteristics evolve. Kernel reuse strategies enable repurposing of computational resources across diverse application domains, with systematic refactoring approaches achieving 50-70% code reusability between related GPU applications (Sarkar, S. *et al.*, 2012). High-performance units initially deployed for training large language models are subsequently repurposed for inference tasks, development environments, or educational applications where reduced computational capacity remains

Publisher: SARC Publisher

acceptable. GPU datacenter scheduling taxonomies identify multiple optimization dimensions, including fairness, efficiency, and resource isolation, with multi-objective scheduling algorithms balancing competing performance requirements across heterogeneous workload portfolios (Gao, W. *et al.*, 2022). Lifecycle extension techniques focus on maintaining hardware performance through proactive maintenance, thermal management optimization, and targeted upgrades, restoring or enhancing GPU capabilities. Kernel optimization and refactoring techniques offer software-driven solutions for prolonging hardware lifespan, facilitating ongoing performance enhancements without the need for hardware replacement cycles (Sarkar, S. *et al.*, 2012). Memory module replacement, cooling system upgrades, firmware optimization, and careful computational workload management minimize wear on critical components while maximizing resource utilization efficiency. Organizations implementing comprehensive lifecycle extension programs leverage advanced scheduling techniques, achieving 35-55% improvement in hardware longevity through intelligent workload distribution and resource allocation strategies (Gao, W. *et al.*, 2022).

Sustainable Workload Management and Optimization Techniques

Sustainable workload management requires sophisticated approaches balancing computational efficiency with hardware preservation and environmental responsibility. Traditional workload schedulers primarily optimize for performance metrics, including job completion time and throughput. At the same time, sustainable approaches must incorporate additional factors, including energy consumption, hardware wear patterns, and long-term resource availability. Multi-objective optimization challenges require advanced algorithms capable of navigating complex trade-offs between immediate performance requirements and sustainable resource utilization. Energy-efficient GPU workload management in heterogeneous cloud environments demonstrates potential for achieving 30-45% energy savings while maintaining 85-90% computational performance through intelligent resource allocation and dynamic scaling techniques (Brown, I. *et al.*, 2025). Energy-aware scheduling algorithms represent a fundamental shift from performance-centric approaches toward holistic resource management, considering power consumption, thermal generation, and cooling requirements. Heterogeneous cloud environments present unique challenges requiring adaptive scheduling strategies capable of managing diverse GPU architectures with varying power characteristics and computational capabilities (Brown, I. *et al.*, 2025). Advanced energy optimization techniques demonstrate the capability to reduce operational power consumption through intelligent workload distribution across GPU clusters with different energy efficiency profiles. GPU performance optimization strategies focus on reducing computational waste through systematic identification and elimination of inefficient processing patterns, achieving 25-40% improvement in resource utilization rates while minimizing environmental impact (Arc Compute, 2024).

Thermal management optimization plays a crucial role in extending hardware lifespan and maintaining performance consistency over extended operational periods. Cloud-based GPU workload management systems incorporate real-time thermal monitoring, enabling proactive temperature control across heterogeneous hardware configurations with diverse cooling requirements (Brown, I. *et al.*, 2025). Advanced thermal modeling techniques predict temperature distributions across GPU clusters, enabling workload placement adjustments that prevent thermal stress accumulation. GPU performance optimization methodologies emphasize waste reduction through systematic analysis of computational bottlenecks and inefficient resource allocation patterns, resulting in enhanced operational efficiency and reduced thermal generation (Arc Compute, 2024). Quality of service management in sustainable AI clusters requires balancing performance guarantees with resource conservation goals. Heterogeneous cloud environments necessitate adaptive quality frameworks capable of dynamically adjusting computational precision levels based on available GPU resources and energy constraints (Brown, I. *et al.*, 2025). Model complexity optimization and adaptive output frequency adjustment enable organizations to maintain service levels during peak demand periods while operating in more efficient modes during normal operations. AI companies implementing comprehensive GPU performance optimization strategies report significant improvements in computational efficiency through systematic waste reduction methodologies, achieving enhanced resource utilization while maintaining acceptable service quality standards (Arc Compute, 2024). Flexible quality management systems contribute to both

Publisher: SARC Publisher

operational cost reduction and environmental sustainability through optimized computational resource demands across diverse hardware architectures.

Table 2: Environmental Impact Reduction through Intelligent Workload Management (Brown, I. *et al.*, 2025; Arc Compute, 2024)

Sustainability Metric	Baseline	Optimized Range	Optimization Focus
Energy Consumption	100	55-70% of baseline	Power Management
Computational Performance	100	85-90% of retention	Service Quality
Resource Utilization Efficiency	100	90% improvement	Waste Reduction
Thermal Generation	100	75-85% of baseline	Cooling Requirements
Quality of Service	100	90-95% maintenance	Performance Guarantee
Operational Cost	100	70-80% of baseline	Economic Efficiency

Cost-Benefit Analysis and Environmental Impact Assessment

A comprehensive cost-benefit analysis of GPU reuse strategies requires examination of both direct financial impacts and broader environmental considerations, translating to long-term economic benefits. Direct cost savings include reduced hardware acquisition expenses, lower operational energy consumption, decreased cooling requirements, and minimized waste disposal fees. Sustainable AI implementation presents significant environmental challenges alongside economic opportunities, requiring careful balance between computational performance and ecological responsibility (Wu, C. J. *et al.*, 2021). Organizations implementing systematic GPU reuse programs demonstrate potential for substantial operational improvements while addressing environmental implications associated with intensive computational workloads. Analysis reveals opportunities for achieving meaningful cost reductions through strategic hardware lifecycle management, with particular emphasis on minimizing environmental impact during AI system deployment and operation (Wu, C. J. *et al.*, 2021). Environmental impact assessment must consider the full lifecycle of GPU hardware, from manufacturing emissions to end-of-life disposal considerations. Sustainable AI research indicates substantial environmental implications arising from computational infrastructure requirements, with hardware manufacturing contributing significantly to the overall carbon footprint (Wu, C. J. *et al.*, 2021). Modern AI applications demand extensive computational resources, creating challenges for environmental sustainability across diverse application domains. Advanced lifecycle management approaches demonstrate potential for reducing environmental burden through optimized resource utilization and extended hardware operational periods. Power consumption optimization strategies enable organizations to achieve meaningful reductions in carbon emissions while maintaining computational effectiveness, with particular benefits observed in large-scale deployment scenarios (Degot, C. *et al.*, 2021). Return on investment calculations for GPU reuse initiatives must account for implementation costs, including staff training, monitoring system development, and process redesign efforts. Organizations with mature reuse programs achieve favorable economic outcomes through a systematic approach to hardware lifecycle management, with emphasis on maximizing value extraction from existing computational resources (Degot, C. *et al.*, 2021). Economic benefits extend beyond direct cost savings to include improved budget predictability through reduced hardware refresh frequency, supply chain risk mitigation, and enhanced corporate sustainability profiles providing competitive advantages in environmentally conscious markets. AI-driven optimization strategies contribute to both operational efficiency improvements and environmental impact reduction, creating dual value propositions for organizations seeking sustainable computational solutions (Degot, C. *et al.*, 2021). Risk assessment for GPU reuse strategies must address potential performance degradation, reliability concerns, and operational complexity, offsetting some benefits. Careful monitoring and progressive workload assignment mitigate most risks while maintaining operational efficiency through systematic hardware health assessment protocols. Organizations report success in implementing reuse programs when combined with comprehensive risk management frameworks addressing environmental implications and sustainability challenges (Wu, C. J. *et al.*, 2021). Performance monitoring systems enable proactive management of hardware reliability while supporting environmental objectives through extended operational lifecycles. Strategic implementation of sustainable AI practices requires balancing computational requirements with environmental responsibility, achieving

Publisher: SARC Publisher

optimal outcomes through an integrated approach to resourcemanagement and ecological stewardship (Degot, C. *et al.*, 2021).

Table 3: Lifecycle Environmental Benefits of GPU Reuse and Optimization Strategies(Wu, C. J. *et al.*, 2021; Degot, C. *et al.*, 2021)

Environmental Assessment Factor	Traditional Approach	Sustainable Approach	Impact Category	Measurement
Manufacturing Emissions Amortization	2-3 years	6-8 years	Carbon Footprint	Lifecycle Extension
Operational Carbon Reduction	Baseline	30-45% reduction	Emission Reduction	Environmental Benefit
Electronic Waste Generation	High	Minimized	Waste Management	Resource Conservation
Resource Utilization Efficiency	15-30%	30-50%	Resource Optimization	Efficiency Improvement
Power Grid Burden	High	Reduced	Infrastructure Impact	System Sustainability
Lifecycle Environmental Impact	Maximum	Optimized	Overall Assessment	Holistic Benefit

CONCLUSION

The imperative for sustainable GPU management emerges as a critical solution to address mounting environmental and economic challenges within artificial intelligence infrastructure deployment. Manufacturing processes for modern graphics processors generate substantial carbon emissions before operational deployment, emphasizing the necessity of extending hardware lifecycles to maximize environmental return on investment. Current utilization patterns demonstrate significant inefficiencies, with computational resources remaining underutilized while continuing to consume power and thermal heat. Systematic implementation of reuse strategies offers viable pathways for achieving substantial reductions in carbon footprints while simultaneously improving organizational return on investment through prolonged hardware operational periods. Advanced scheduling algorithms and intelligent resource allocation frameworks enable maintenance of computational performance standards while optimizing energy consumption patterns and thermal management protocols. Integration of kernel optimization methodologies and software-level performance enhancements provides additional opportunities for extending hardware utility without requiring complete system replacements. Economic benefits of sustainable GPU management extend beyond direct cost savings to include improved budget predictability, reduced supply chain dependencies, and enhanced corporate sustainability profiles, providing competitive advantages in environmentally conscious markets. Organizations adopting comprehensive reuse frameworks demonstrate potential for meaningful environmental impact reduction while maintaining operational efficiency requirements. The convergence of environmental responsibility and economic pragmatism creates compelling justification for transitioning from traditional hardware replacement cycles toward sustainable resource management practices. Success requires systematic approaches encompassing hardware assessment protocols, adaptive workload management systems, and lifecycle extension techniques tailored to organizational computational requirements.

REFERENCES

1. Iftikhar, S. & Davy, S. "Reducing carbon footprint in ai: A framework for sustainable training of large language models." *Proceedings of the Future Technologies Conference*. Cham: Springer Nature Switzerland, (2024).
2. Klizo Solutions Pvt. Ltd, "Carbon Footprint of AI: The Environmental Cost of Training LLMs," *Medium*, 26 September (2024). Available:https://klizosolutions.medium.com/carbon-footprint-of-ai-the-environmental-cost-of-training-llms-13b3688d7b7e
3. Marwala, T. "Rethinking Tech and Why GPUs Are Not the Future of AI Training," *United Nations University (UNU)*, 2 April (2025). Available:https://unu.edu/article/rethinking-tech-and-why-gpus-are-not-future-ai-training
4. Bridge, B. "Impact of AI Performance Efficiency on Long-Term GPU Demand: The Case of DeepSeek AI," Medium, 30 January (2025). Available:https://bytebridge.medium.com/imp

Publisher: SARC Publisher

act-of-ai-performance-efficiency-on-long-term-gpu-demand-the-case-of-deepseek-ai-7d5f607e9b9c

5. Sarkar, S., Mitra, S., & Srinivasan, A. "Reuse and refactoring of GPU kernels to design complex applications." *2012 IEEE 10th International Symposium on Parallel and Distributed Processing with Applications*. IEEE, (2012).
6. Gao, W., Hu, Q., Ye, Z., Sun, P., Wang, X., Luo, Y., ... & Wen, Y. "Deep learning workload scheduling in GPU datacenters: Taxonomy, challenges and vision." *arXiv preprint arXiv:2205.11913* (2022).
7. Brown, I. *et al.*, "Energy-Efficient GPU Workload Management in Heterogeneous Cloud Environments," ResearchGate, March (2025). Available: https://www.researchgate.net/publication/391904908_Energy-Efficient_GPU_Workload_Management_in_Heterogeneous_Cloud_Environments
8. Arc Compute, "Optimizing GPU Performance for Developing AI-Strategies to Reduce Waste and Enhance Efficiency," 12 March (2024). Available: https://www.arccompute.io/arc-blog/optimizing-gpu-performance-for-ai-companies-strategies-to-reduce-waste-and-enhance-efficiency
9. Wu, C. J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., ... & Hazelwood, K. "Sustainable AI: Environmental Implications." *Challenges and Opportunities* (2021).
10. Degot, C., Duranton, S., Frédeau, M., & Hutchinson, R. "Reduce Carbon and Costs with the Power of AI." *Boston Consulting Group* 26 (2021).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Gupta, S. "Reusing GPUs to Reduce Carbon Footprints and Costs in AI Clusters." *Sarcouncil Journal of Multidisciplinary 5.7* (2025): pp 404-410.

Publisher: SARC Publisher