

Geographical Server Relocation: Opportunities and Challenges

Yejia Liu[†]
UC Riverside
yliu807@ucr.edu

Daniel Wong
UC Riverside
danwong@ucr.edu

Pengfei Li[†]
UC Riverside
pli081@ucr.edu

Shaolei Ren*
UC Riverside
shaolei@ucr.edu

ABSTRACT

The enormous growth of AI computing has led to a surging demand for electricity. To stem the resulting energy cost and environmental impact, this paper explores opportunities enabled by the increasing hardware heterogeneity and introduces the concept of Geographical Server Relocation (GSR). Specifically, GSR *physically* balances the available AI servers across geographically distributed data centers subject to AI computing demand and power capacity constraints in each location. The key idea of GSR is to relocate older and less energy-efficient servers to regions with more renewables, better water efficiencies and/or lower electricity prices. Our case study demonstrates that, even with modest flexibility of relocation, GSR can substantially reduce the total operational environmental footprints and operation costs of AI computing. We conclude this paper by discussing major challenges of GSR, including service migration, software management, and algorithms.

KEYWORDS

AI, geographical server relocation, sustainability, cost

1 INTRODUCTION

As we embark on the era of artificial intelligence (AI) characterized by the widespread adoption of advanced models such as ChatGPT and MidJourney, the significant energy consumption involved in training, inference, and fine-tuning these AI models is increasingly worrisome. For example, training a single large language model takes millions of GPU hours and consumes electricity in the order of thousands of megawatt hours [2, 29].

Consequently, concerns regarding the environmental footprints and energy costs of data centers housing AI servers have garnered significant attention. A recent estimate conducted by the International Energy Agency projects a sharp increase in the global AI energy demand, reaching at least ten times the current level and exceeding the annual electricity consumption of a small country like Belgium by 2026 [11]. In light of the surging AI demand, there has been a pressing need to implement cost-efficient and eco-friendly solutions to ensure a sustainable future for AI development.

Numerous strategies have been pursued to address the huge electricity cost and environmental impacts of AI computing. For example, reducing AI model sizes through model compression, speeding up AI training and inference, and/or adopting GPUs and

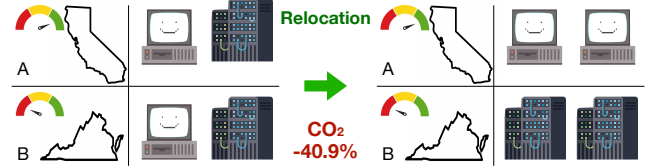


Figure 1: Illustration of potential carbon emission reduction through GSR by relocating energy-efficient servers to Virginia and energy-inefficient servers to California.

purpose-built accelerators [8, 16, 17, 21] can yield substantial energy efficiency improvement. On the other hand, different locations exhibit significant degrees of geographical heterogeneities in terms of their electricity prices, average carbon intensities of the grids, and/or the climate conditions that affect the water efficiencies. Thus, another line of efforts being extensively studied involves leveraging the spatial and temporal flexibility inherent in AI computing workloads [10]. This entails dynamically adjusting the location and timing of AI computing to better align with periods and locations where low-carbon and/or low-cost energy sources are available [3, 15]. The emergence of third-party energy information services, such as offering real-time data on energy’s carbon intensity at high resolutions, has lowered the barrier for this approach and made it more viable [1, 7, 9]. Importantly, such AI workload shifting across different geographical locations has been increasingly adopted by major technology companies as an effective enabler for sustainable computing [22].

While the potential of geographically shifting AI computing workloads has been well-recognized, another complementary knob — *physically* moving AI computing servers around geographically distributed data centers — has remained largely over-looked for sustainability and cost-saving. We refer to this approach as *Geographical Server Relocation* (GSR).

The rationale that motivates our pursuit of GSR comes from the increasing hardware heterogeneity. Concretely, despite the development of more powerful and energy-efficient servers, the high cost remains a barrier, making it impractical or financially challenging to replace all the servers with the latest, expensive, and more energy-efficient AI hardware at once. Instead, partial refreshment is more common in the data center upgrade lifecycle [23]. This practice has also been reinforced by the increasing emphasis on reducing the servers’ embodied environmental footprint during the manufacturing process [6]. As a result, today’s AI data centers often feature heterogeneous architecture compositions, comprising a mix of older and newer AI servers. Therefore, the total environmental footprints and operational costs of AI computing can be reduced

[†]Yejia Liu and Pengfei Li contributed equally.

*Corresponding author: Shaolei Ren (shaolei@ucr.edu).

by strategically relocating older and less energy-efficient servers to regions where renewable energy sources are more abundant, water efficiency is higher, and/or electric prices are lower.

We show an illustrative example in Figure 1 as a thought experiment. Consider two AI data center locations, such as California and Virginia in Figure 1, labeled as A and B, respectively. Virginia's carbon intensity is roughly three times higher than California's, based on their average carbon intensity (around 130g/kWh for California vs. 369g/kWh for Virginia) in April 2024 according to Electricity Maps [18]. There are two types of AI servers: normalized performance per watt is 10 for newer servers and 1 for older ones. Suppose that we have two units of AI workloads, equally split between the two types of servers. In other words, the normalized quantities of older servers and newer servers are 1 and 1/10, respectively, in each data center before relocation.

- *Before relocation:* The total normalized carbon emissions at both locations is $(1/1 + 1/10) * 1 + (1/1 + 1/10) * 3 = 4.4$.

- *After relocation:* In this case, we relocate less energy-efficient older servers from Virginia to California which has a lower carbon intensity. Meanwhile, to meet the pre-relocation AI computing demand at each location, we relocate the newer servers from California to Virginia. Thus, the total normalized carbon emissions of these two locations become $(1/1 + 1/1) * 1 + (1/10 + 1/10) * 3 = 2.6$, resulting in $\sim 40.9\%$ reduction in operational carbon emission compared to the pre-relocation level.

Despite the oversimplification of many practical considerations, the illustrative example above demonstrates a clear potential of GSR to reduce AI's surging environmental footprint in light of the increasing hardware heterogeneity. In this paper, we further formalize the problem of GSR and conduct a case study to highlight the potential reductions in carbon emissions, water consumption, and electricity costs that GSR may achieve empirically. Nonetheless, compared to shifting AI computing workloads around different locations, GSR presents additional challenges in terms of service migration, software management, and algorithms, among others. Thus, to offer a more balanced view, we will highlight these challenges in this paper, which we hope can shape some interesting research directions for the community to realize the full potential of GSR for sustainability and cost saving.

2 OPPORTUNITIES FOR GSR

In this section, we present the emerging opportunities for GSR enabled by the hardware and geographical heterogeneities.

2.1 Hardware Heterogeneity

Most AI computing workloads run on GPUs nowadays, whose energy efficiency has increased dramatically in recent years due to design optimization and architectural advances [20]. Figure 2 shows the normalized performance per watt of data center-grade GPUs released by Nvidia from 2014 to 2023, demonstrating a more than 10x improvement in terms of GFLOPS per watt.

Despite the significant improvement, the high upfront costs compounded by supply chain constraints make it impractical or financially challenging to replace all the servers with the latest, expensive, and more energy-efficient AI hardware at once. As a result, it is a common practice for AI developers to partially refresh

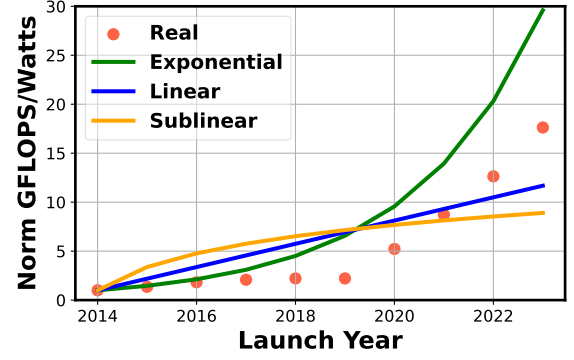


Figure 2: Normalized ratio of performance (GFLOPS) to power consumption (Watts) for data center-grade GPUs over the past 10 years (2014-2023) [20]. The manufacturer-reported data points are plotted in dots and labeled as “real”. We also offer three different synthetic curves (i.e., “exponential”, “linear”, and “sublinear”) for hypothetical studies.

and upgrade their AI server fleet, resulting in a mixture of new and old AI servers [6, 23]. Further, the improved server reliability and increasing emphasis on reducing AI servers' embodied environmental footprint during the manufacturing process has propelled a growing trend of keeping servers for a longer lifespan before retirement [6]. More recently, composing servers using retired components (e.g., DRAMs and CPUs) has also been proven effective for cutting servers' lifecycle carbon footprints.

These practices have led to a significant AI hardware heterogeneity in terms of the performance per watt in many data centers.

2.2 Geographical Heterogeneity

To serve users worldwide, AI data centers are located in different regions, which also exhibit significant geographical heterogeneities.

- *Electricity price:* There is a significant spatial variation of electricity prices across different states and countries [3]. For example, country-wide electricity prices can differ by more than 10x throughout the world [24].

- *Carbon intensity:* The regional differences in energy sources for electricity generation naturally result in significant disparities in carbon intensities for each kWh of electricity consumption [18]. Even though technology companies have increasingly adopted carbon-free energy for powering their global data centers, such regional differences still persist. For example, 97% of the energy usage by Google's data center in Finland is carbon-free, whereas this number drops to 4-18% for its data centers in Asia [6].

- *Water efficiency:* In addition to carbon emissions, AI computing also has a significant water footprint, which has emerged as a hidden sustainability roadblock [14]. Water efficiency in terms of water consumption per kWh of IT energy usage, a.k.a., water usage effectiveness (WUE), also varies significantly across different locations (e.g., by more than 20x across Microsoft's global data center locations). Importantly, a data center with better carbon efficiency may have worse water efficiency [14]. This necessitates

AI computing's water consumption as a separate sustainability metric to address.

2.3 Opportunities Enabled by Heterogeneity

Many AI training and inference tasks can run a diverse set of GPUs without necessarily having to use a specific type of GPU. That is, to meet the same AI computing demand, there exist an increasingly wider set of AI servers, each with different performances per watt. On the other hand, geographical heterogeneities mean that even with the same utilization, the same server can have very different energy costs and environmental footprints if put in different data centers. As such, where to place the available AI servers to meet the demand in each data center becomes an important question.

This motivates our pursuit of GSR to tap into the potential opportunities enabled by hardware and geographical heterogeneities for sustainability and cost saving. For example, as illustrated in Figure 1, GSR can relocate older and less energy-efficient servers to regions with more renewables, better water efficiencies and/or lower electricity prices subject to AI computing demand and power capacity constraints in each data center.

2.4 Problem Formulation

Suppose that there are N data centers and up to M types of AI hardware/servers in each data center with different performances per watt. We denote the default/current and the new configurations of AI servers as $\mathbf{y} = \{y_{i,j} | i \in N, j \in M\}$ and $\mathbf{x} = \{x_{i,j} | i \in N, j \in M\}$, representing the pre-GSR and the post-GSR quantities of type- j servers in each data center i , respectively.

The computational capacity of type- j AI server in data center i is defined as $f_j(x_{i,j})$, which can be measured in terms of GFLOPS or other metrics that AI developers use for capacity planning purposes. Given the average utilization, the energy consumption of type- j AI servers in data center i is denoted as $e_j(y_{i,j})$. Similarly, we denote its power consumption as $p_j(y_{i,j})$, which indicates the required power capacity to support the server deployment.

Operational cost. The operational cost (energy cost, carbon footprint, water consumption, or a combination of them) is proportional to the total energy consumption of all the AI servers in each data center. Thus, we use a linear function $H_{i,c}(\sum_{j=1}^M e_j(x_{i,j})) = q_i \sum_{j=1}^M e_j(x_{i,j})$ to represent the cost of data center i , where q_i is the average electricity price, carbon intensity, WUE, or a combination. The coefficient q_i also absorbs the average power usage effectiveness (PUE) to account for non-IT energy overheads if applicable.

Relocation cost. GSR introduces server relocation costs, such as the shipping costs and carbon emission overheads due to logistics (which are usually small compared to servers' operational emissions in the lifecycle). Here, to capture the relocation costs, we use the difference $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ between pre-GSR configuration $\mathbf{x}$ and post-GSR server configuration $\mathbf{y}$ as a proxy measure. Given two different configurations $\mathbf{x}$ and $\mathbf{y}$, the actual relocation cost can be obtained by optimizing the server relocation schedule (e.g., where and which servers in a data center should be relocated).

Constraints. We introduce $\rho_i \in [0, 1]$ to denote the fraction of AI computing capacity that needs to be retained in data center i . When $\rho_i = 1$, we must ensure that the post-GSR and pre-GSR computational capacities of AI servers are the same; when there is

maximum flexibility at $\rho_i = 0$, we can even shut down data center i entirely and relocate all the AI servers to elsewhere, which can apply to AI developers who rent data center spaces from third-party colocation providers (e.g., Equinix).

Additionally, we use $\gamma_i \geq 1$ to denote the extra power capacity available normalized by the pre-GSR usage level in data center i . Typically, data center operators reserve extra capacity to absorb additional loads and accommodate for future growth. If an AI developer rents power capacity from a third-party provider, it can have even more flexibility (i.e., a larger γ_i). For notational conveniences, we also absorb physical space constraints into γ_i for data center i .

Next, we formalize the problem of GSR as follows:

$$\min_{\mathbf{x}} \sum_{i=1}^N H_{i,c} \left(\sum_{j=1}^M e_j(x_{i,j}) \right) + \mu_d \cdot d(\mathbf{x}, \mathbf{y}) \quad (1a)$$

$$s.t. \quad \sum_{j=1}^M f_j(x_{i,j}) \geq \rho_i \sum_{j=1}^M f_j(y_{i,j}), \quad \forall i \in [1, N] \quad (1b)$$

$$\sum_{j=1}^M p_j(x_{i,j}) \leq \gamma_i \sum_{j=1}^M p_j(y_{i,j}), \quad \forall i \in [1, N] \quad (1c)$$

$$\sum_{i=1}^N x_{i,j} = \sum_{i=1}^N y_{i,j}, \quad \forall j \in [1, M] \quad (1d)$$

The objective (1a) is a weighted sum of the operational cost and the relocation cost, with the weight hyperparameter $\mu_d \geq 0$ denoting the unit relocation cost. The constraint (1b) specifies the minimum post-GSR AI computing capacity relative to the pre-GSR level, the constraint (1c) specifies the power capacity constraint, and the constraint (1d) means that GSR does not retire any available AI servers (which is a separate decision beyond the scope of GSR).

Remark. GSR only relocates *existing* servers that have already been purchased (or refurbished if re-built from older servers [28]); it does not decide whether or not to buy new AI hardware, which can be an interesting future study but is beyond the current scope of GSR. As such, all the potential benefits of GSR shown in this paper lie in the operational costs and environmental footprints, rather than the capital expenses and embodied footprints.

3 CASE STUDY

In this section, we conduct a case study to evaluate the potential empirical effectiveness of GSR under a synthetic setting based on the reported GPU energy efficiency over the last 10 years.

3.1 Experimental Setup

We examine a set of 10 geographically-distributed data centers based Microsoft's current data center sites [19]. This set comprises four situated in the United States (Virginia, Georgia, Texas, and Nevada), four in Europe (Belgium, the Netherlands, Germany, and Denmark), and two in Asia (Singapore and Japan).

3.1.1 Datasets.

Data center-grade GPU energy efficiency. We utilize the general information regarding Nvidia GPUs tailored for data center usage over the past decade (2014-2023) as documented by [20]. Specifically,

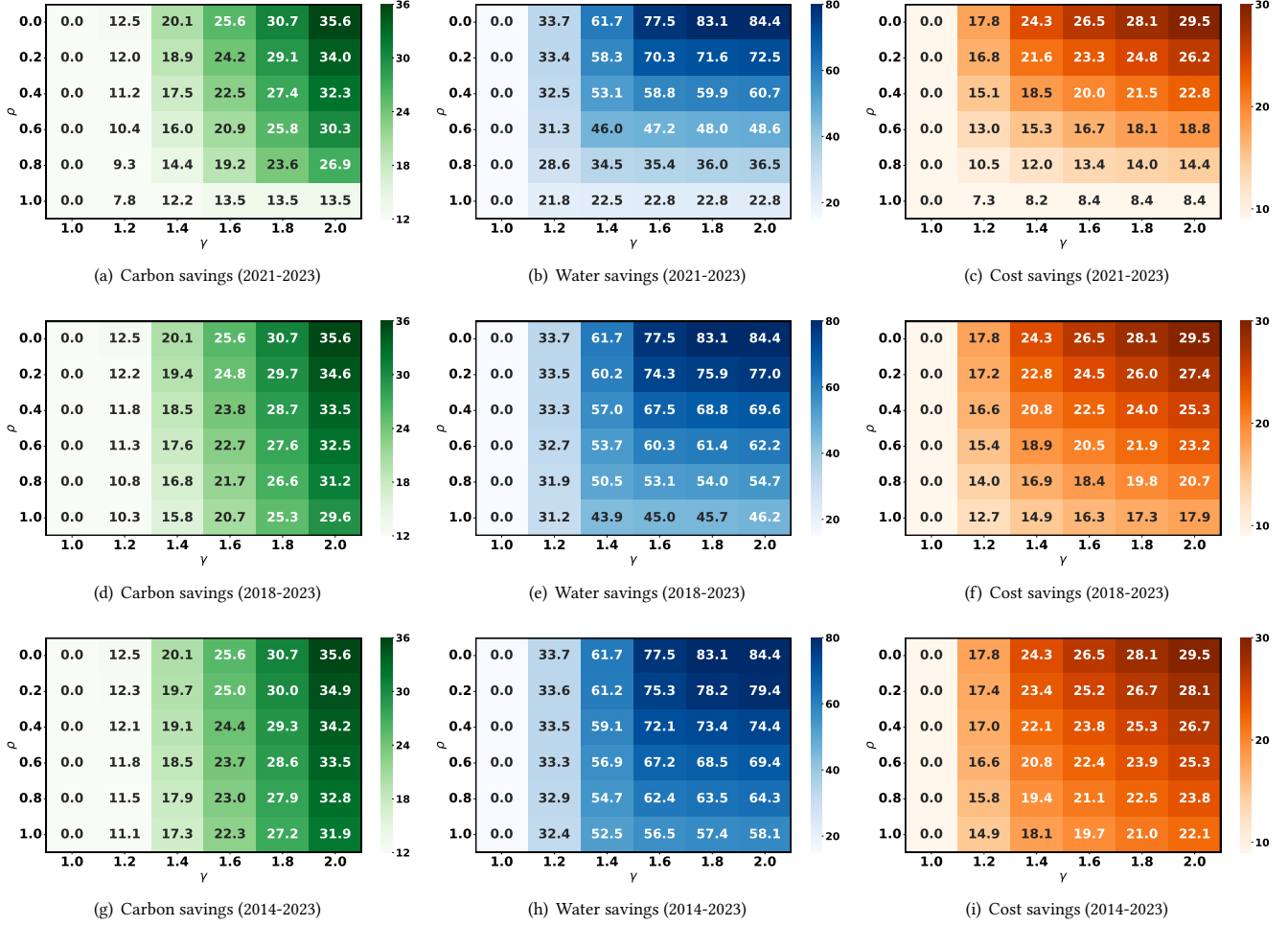


Figure 3: Carbon emission, water consumption, and electricity cost savings under different γ and ρ assessed by the manufacturer-reported GPU performance-to-power data.

we consider the performance using the provided GFLOPS (Single-precision) to measure the computational capacity f_j , and thermal design power (TDP) in Watts to gauge the energy consumption e_j of each type- j AI server. To ensure comparability across the ten-year timeframe, we normalize performance-to-power ratios by establishing 2014 as the baseline, setting its value to 1. We show the normalized performance (GFLOPS) per watt in Figure 2. Based on the real manufacturer-reported data represented by the dotted line, we can observe a clear trend of rapid increases in performance per watt as the year progresses.

Electricity price, WUE, PUE, and carbon intensity. We obtain yearly average electricity price for each data center location in Europe and Asia from [12]. For the U.S. data centers, we collect the electricity prices from their respective ISOs as documented in [25]. In terms of environmental footprint minimization, we primarily focus on operational carbon emission and water consumption [14, 29]. Specifically, we use the on-site cooling WUE for these 10 data centers reported by [14]. As for the carbon intensity, we gather yearly

average data across these 10 data center locations for the most recent years from [18]. We use the annualized average PUE for each data center based on Microsoft’s most recent disclosure [19].

3.1.2 Default configuration and metric. Because of competitive reasons, there is no precise information in the public domain regarding the current configuration of AI servers in each data center. Thus, for the pre-GSR setting, we assume that the AI servers are uniformly distributed in terms of their power consumption. That is, before GSR for the latest three years (2021, 2022, 2023), we assume the same amount of power consumption by AI servers purchased from each year in each data center. We will also consider other settings such as non-uniform pre-GSR configurations and longer-time scales (see Appendix A).

We evaluate the effectiveness of GSR by quantifying the percentage of savings in operational electricity costs as well as reductions in the environmental footprint before and after GSR.

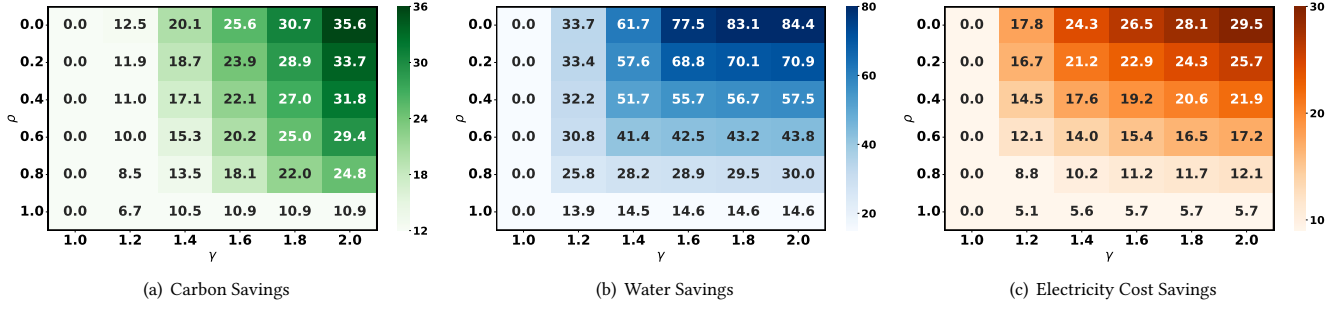


Figure 4: Carbon emission, water consumption, and electricity cost savings under different γ and ρ assessed by the manufacturer-reported GPU performance-to-power data over the last 3 years (2021, 2022, 2023). The idle power of the servers is included.

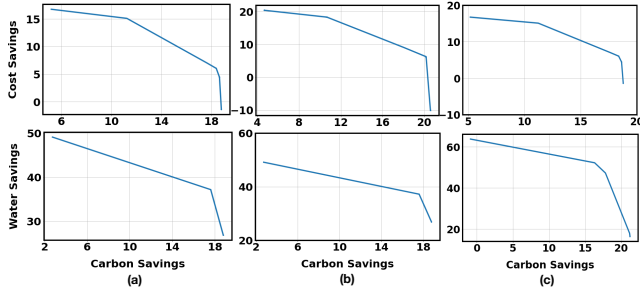


Figure 5: Carbon footprint, water consumption, and electricity cost saving tradeoffs under the manufacturer-reported GPU data over: (a) the latest 3 years (2021, 2022, 2023); (b) the latest 6 years (2018-2023); (c) the latest 10 years (2014-2023).

3.2 Numerical Results

Our empirical results demonstrate that, even with modest flexibility of relocation, GSR can dramatically reduce the total environmental footprints and operational costs compared to the pre-GSR level.

3.2.1 Results for manufacturer-reported data. In Figure 3, we present the reduction in operational carbon footprint, water consumption, as well as the electricity cost savings before and after GSR under various combinations of $\rho \in [0, 1]$ and $\gamma \geq 1$ values, considering different time frames of GPU performance per watt data. Specifically, we only minimize the individual cost metric (e.g., electricity cost) without considering the other metrics or relocation costs (i.e., setting $\mu_d = 0$). Thus, the values in Figure 3 represent the maximum savings for the respective metrics under different ρ and γ .

As $\rho > 0$ decreases, GSR can potentially relocate more AI servers since less AI computing demand needs to be processed in the same data center after GSR. Likewise, with a larger $\gamma \geq 1$, the extra power capacity available for GSR is larger, which enhances the flexibility of GSR. Therefore, we observe that as the flexibility of GSR increases, the potential saving becomes significantly larger. Importantly, with a modest flexibility (e.g., $\rho = 0.5$ and $\gamma = 1.5$), GSR can roughly yield 20%, 50+%, and 20% savings in terms of the operational carbon footprint, water consumption, and electricity cost, respectively.

3.2.2 The impact of idle power. Our results in Figure 3 focus on the dynamic GPU power only. In practice, however, active servers also have idle power even when they are not processing any AI workloads. Thus, we now investigate the impact of such idle power on GSR. Specifically, each modern GPU-based AI servers typically houses 4-8 GPUs. Considering the peak power consumption of 130W for CPU (e.g., Intel Xeon W-2125 processor) and the typical power draw of around 5W for a 16GB DDR4 RAM, we add an effective amortized idle power of 30W to each GPU when calculating the performance-to-power value.

By using the same pre-GSR configuration as in Figure 3, we show the cost savings for carbon, water, and electricity while accounting for idle power in Figure 4. Despite slightly decreased savings, the carbon, water and electricity cost reductions achieved by GSR are highly similar to those in Figure 3, emphasizing that the primary driver for savings comes from the spatial and server heterogeneity. For example, relocating servers to regions with a lower carbon footprint can significantly decrease carbon emissions, even with some idle power consumption added to the servers. Similarly, the spatial heterogeneity in water efficiency and electricity prices plays a crucial role in savings for water and electricity cost, respectively.

3.2.3 Carbon vs. water (electricity cost) tradeoffs. Carbon efficiency, water efficiency, and electricity cost efficiency are three important, but often conflicting, objectives [13]. For example, California has a higher water consumption rate due to its drier and hotter climate than Virginia, but its carbon intensity for electricity generation is much lower than Virginia's. Likewise, despite the cleaner energy sources for electricity generation, the electricity price in California is higher than that in many other U.S. states.

Thus, we show the tradeoff between carbon emission reduction vs. water consumption reduction, and carbon emission reduction vs. electricity cost reduction. Specifically, we minimize the weighted sum of carbon emission and water consumption/electricity cost, and vary the weight. The results are shown in Figure 5, where we set $\rho = 0.5$, $\gamma = 1.5$ and $\mu = 0$, based on the manufacturer-reported GPU performance-to-power data spanning the latest 3 years (2021, 2022, 2023), the latest 6 years (2018-2023), and the latest 10 years (2014-2023), respectively.

While different metrics may not be perfectly aligned, GSR can still simultaneously reduce AI's carbon emission, water consumption,

and electricity cost, which may not be achievable by geographical load balancing alone that only shifts workloads across different data centers [13]. Interestingly, when aggressively minimizing carbon emissions, we may end up with a higher electricity cost in some cases, which corroborates with the prior finding that carbon-efficient locations may not be cost-effective [4].

Due to space limitations, we defer additional results to Appendix A, including the impact of relocation costs and the results for synthetic trends of GPU performance per watt.

4 CHALLENGES FOR GSR

While GSR could potentially reduce the total environmental footprints and operational costs, it also creates new challenges.

4.1 Services Impacted by GSR

By consolidating the available hardware as a resource pool, modern cluster management can easily handle individual server replacement/installation without affecting the running services [27], and can even handle unexpected data center-wide failures by temporarily relocating all the impacted workloads to other data centers [26]. The physical relocation process in GSR requires unplugging move-out servers and plugging move-in servers, and also requires spare server capacity (which typically exists to handle workload variations and growth) to temporarily process the impacted workloads. Thus, GSR can be viewed as a planned global-level maintenance event, which presents additional systems challenges. Alternatively, one can optimize the server relocation schedule and execute the relocation decision for one data center after another to minimize the impacted AI servers as well as the hosted workloads.

4.2 Software and Workloads

GSR may present several challenges for software that is highly tuned for specific hardware features. For example, to maintain performance in virtualized environments, software runtimes may depend on hardware isolation mechanisms, such as cache partitioning with Intel Cache Allocation Technology (CAT). Similarly, certain software may rely on vectorization for performance using AVX-512, while other processors may only support AVX2 extensions. These variations in hardware features may exist in only certain families of CPUs, which can present challenges to the software when running on relocated servers.

Besides software challenges, workloads that run on the relocated hardware may also need to adapt due to differences in cache/memory hierarchy, processor core type (performance cores vs efficiency cores), and parallelism available in the processors. This challenge is already commonly experienced in high-performance computing (HPC) systems. For example, thousands of man-hours are spent porting workloads from one HPC system to a new generation of HPC systems. Due to this, there is a strong focus on portability to achieve high performance across diverse hardware. To better support GSR, this focus on software portability needs to be adopted in cloud environments.

Workload changes may also need to adapt to changing hardware features. For example, older GPUs may not have hardware features such as tensor cores, or support for system-wide atomics. The absence of tensor cores would require algorithms to fall

back to traditional arithmetic units for computation. System-wide atomics greatly simplifies the implementation of distributed GPU algorithms. The absence of system-wide atomics would require an increased amount of synchronization, leading to programmer burden and decreased performance.

Nonetheless, these challenges may be less of an issue if the software and/or workloads closely tied to specific hardware features can be relocated together with the associated servers and run elsewhere (i.e., $\rho_i < 1$ in our formulation (1b)).

4.3 Demand and Hardware Uncertainties

When planning for GSR, we need the projected AI resource demand as the input for decision optimization. For example, the parameter ρ governing the fractions of AI demand that cannot be relocated needs to be provided for GSR optimization. Additionally, when the future GPU energy efficiency improves, we might need to relocate some of the AI servers again, potentially resulting in higher movement costs. In other words, we need to solve a sequential decision-making problem with movement costs subject to future uncertainties. This is commonly referred to as smoothed online optimization which penalizes frequent changes in decisions [5], and is known to be challenging even under simplified assumptions. Thus, GSR presents an interesting online optimization problem, which can be of interest to the operational research and optimization community.

5 CONCLUDING REMARKS

In this paper, we explore potential opportunities enabled by the increasing hardware heterogeneity and introduce the novel concept of GSR. By relocating older and less energy-efficient servers to regions with more renewables, better water efficiencies and/or less electricity prices, GSR can substantially reduce the total operational environmental footprints and operation costs of AI computing. We also discuss the major challenges of GSR, including service impacted by GSR, software management, and optimization algorithms.

Being complementary to the well-studied geographic workload balancing, GSR represents an untapped knob that holds a great potential to cut AI's enormous operational energy cost, carbon emissions, and/or water consumption. The challenges of implementing GSR can potentially define future research directions to realize the full potential of GSR.

ACKNOWLEDGEMENT

This work was supported in part by the U.S. National Science Foundation under grants CNS-2007115, CNS-2047521, CCF-2324941 and CCF-2324940.

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. <https://doi.org/10.1145/3575693.3575754>
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya

- Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- [3] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (Today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems* (Boston, MA, USA) (*HotCarbon '23*). Association for Computing Machinery, New York, NY, USA, Article 11, 7 pages. <https://doi.org/10.1145/3604930.3605705>
 - [4] Peter Xiang Gao, Andrew R. Curtis, Bernard Wong, and Srinivasan Keshav. 2012. It's not easy being green. *SIGCOMM Comput. Commun. Rev.* (2012).
 - [5] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. 2019. Beyond online balanced descent: an optimal algorithm for smoothed online optimization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 168, 11 pages.
 - [6] Google. 2023. <https://sustainability.google/reports/>. Environmental Report.
 - [7] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. 2020. Chasing Carbon: The Elusive Environmental Footprint of Computing. *arXiv:2011.02839* [cs.AR]
 - [8] Song Han, Huizi Mao, and William J. Dally. 2016. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. In *ICLR*.
 - [9] Walid A. Hanafy, Qianlin Liang, Noman Bashir, David Irwin, and Prashant Shenoy. 2023. CarbonScaler: Leveraging Cloud Workload Elasticity for Optimizing Carbon-Efficiency. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 7, 3 (Dec. 2023), 1–28. <https://doi.org/10.1145/3626788>
 - [10] Walid A. Hanafy, Qianlin Liang, Noman Bashir, Abel Souza, David Irwin, and Prashant Shenoy. 2024. Going Green for Less Green: Optimizing the Cost of Reducing Cloud Carbon Emissions. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3* (<conf-loc>, <city>La Jolla</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) (*ASPLOS '24*). Association for Computing Machinery, New York, NY, USA, 479–496. <https://doi.org/10.1145/3620666.3651374>
 - [11] International Energy Agency. 2024. Electricity 2024. <https://www.iea.org/reports/electricity-2024> (2024).
 - [12] International Energy Agency (IEA). [n. d.]. Data and statistics. <https://www.iea.org/data-and-statistics>.
 - [13] Mohammad A. Islam, Kishwar Ahmed, Hong Xu, Nguyen H. Tran, Gang Quan, and Shaolei Ren. 2018. Exploiting Spatio-Temporal Diversity for Water Saving in Geo-Distributed Data Centers. *IEEE Transactions on Cloud Computing* 6, 3 (2018), 734–746. <https://doi.org/10.1109/TCC.2016.2535201>
 - [14] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2023. Making AI Less “Thirsty”: Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv 2304.03271* (2023).
 - [15] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. 2024. Towards Environmentally Equitable AI via Geographical Load Balancing. In *e-Energy*.
 - [16] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. *arXiv preprint arXiv:2306.00978* (2023).
 - [17] Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2024. Llm-pruner: On the Structural Pruning of Large Language Models. *Advances in neural information processing systems* 36 (2024).
 - [18] Electricity Maps. [n. d.]. <https://app.electricitymaps.com/>.
 - [19] Microsoft. <https://local.microsoft.com/>. Microsoft in Your Community.
 - [20] Nvidia. 2024. List of Nvidia graphics processing units. https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units
 - [21] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training. *arXiv:2104.10350* [cs.LG]
 - [22] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyu Xiao, Maya Haridasan, Patrick Hung, Nick Care, Saurav Talukdar, Eric Mullen, Kendal Smith, MariEllen Cottman, and Walfredo Cirne. 2023. Carbon-Aware Computing for Datacenters. *IEEE Transactions on Power Systems* 38, 2 (2023), 1270–1280. <https://doi.org/10.1109/TPWRS.2022.3173250>
 - [23] Stefano Sebastio, Kishor S Trivedi, and Javier Alonso. 2018. Characterizing machines lifecycle in google data centers. *Performance Evaluation* 126 (2018), 39–63.
 - [24] Statista. 2023. <https://www.statista.com/statistics/263492/electricity-prices-in-selected-countries/>. Household Electricity Prices Worldwide in September 2023, by Select Country.
 - [25] U.S. Energy Information Administration. [n. d.]. Open data. <https://www.eia.gov/opendata/>.
 - [26] Kaushik Veeraraghavan, Justin Meza, Scott Michelson, Sankaralingam Panneerselvam, Alex Gyori, David Chou, Sonia Margulis, Daniel Obenshain, Shruti Padmanabha, Ashish Shah, Yee Jiun Song, and Tianyin Xu. 2018. Maelstrom: Mitigating Datacenter-level Disasters by Draining Interdependent Traffic Safely and Efficiently. In *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation* (Carlsbad, CA, USA) (*OSDI'18*). USENIX Association, USA, 373–389.
 - [27] Abhishek Verma, Luis Pedrosa, Madhukar R. Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. 2015. Large-scale cluster management at Google with Borg. In *Proceedings of the European Conference on Computer Systems (EuroSys)*. Bordeaux, France.
 - [28] Jaylen Wang, Daniel S. Berger, Fiodar Kazhamiaka, Celine Irvine, Chaojie Zhang, Esha Choukse, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, Ricardo Bianchini, and Akshitha Sriraman. 2024. Designing Cloud Servers for Lower Carbon. In *ISCA*.
 - [29] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable AI: Environmental Implications, Challenges and Opportunities. In *Proceedings of Machine Learning and Systems*, Vol. 4. 795–813.

APPENDIX

A ADDITIONAL NUMERICAL RESULTS

We offer additional numerical results for our case study as follows.

A.1 Impact of μ_d

We show in Figure 6 the reduction in environment footprint and operational cost by varying the weight μ_d for relocation costs to give different penalties for server relocation. We consider modest flexibility by fixing $\rho = 0.5$ and $\gamma = 1.5$. The average hardware configuration change percentage across data centers before and after GSR is computed as $\frac{1}{N} \sum_i (\sum_j |x_{i,j} - y_{i,j}| / \sum_j y_{i,j})$. Specifically, this value inversely correlates with the value of μ , suggesting that a smaller $\mu_d \geq 0$ encourages more server relocation and enables more flexible decisions by GSR. Naturally, from Figure 6, we can see that larger configuration changes also lead to greater savings.

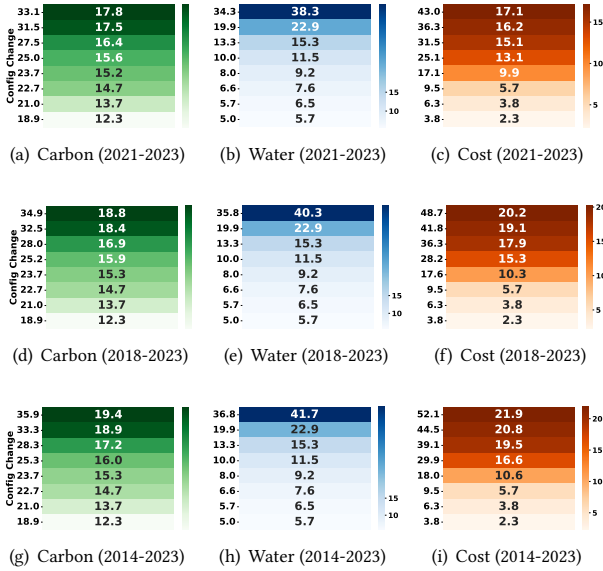


Figure 6: Carbon emission, water consumption, and electricity cost savings for different movement cost weight μ_d in GSR. The y-axis shows the percentage change of server configuration after applying GSR.

A.2 Synthetic Performance-to-Power Trends

In addition to using the manufacturer-reported GPU data to calculate cost savings achieved by GSR, we also incorporate three different synthetic GPU performance-to-power (GFLOPS/Watts) trends to assess the effectiveness of GSR under various setups. Illustrated by the solid-line curves in Figure 2, the exponential increasing rate represents an optimistic estimation, suggesting that the performance per watt of data center-grade GPUs will continue to grow exponentially. In contrast, the linear and sublinear curves project a more neutral or slightly pessimistic estimation, indicating that the performance per watt of GPUs may have already reached a saturation point (perhaps within a few years or at a certain point

in the future). In addition, the sublinear case partially captures the practical observation that the real performance-to-power curve is typically lower than manufacturer-reported values.

In Figure 7, we present the results of environmental and electricity cost savings for these different scenarios, demonstrating that GSR remains effective across these conditions. Specifically, under the same values of ρ and γ , the savings are more prominent in scenarios with exponential GPU performance-to-power trends compared to linear and sublinear trends. It is noteworthy that as ρ and γ reach certain values, the savings under different trends are almost the same, since GSR already relocates most AI servers to the same data center with the best cost efficiency.

A.3 Different Pre-GSR Server Configurations

In our default setting, we consider the case where the AI servers are uniformly distributed in terms of their power consumption. That is, when considering the latest three years (2021, 2022, 2023), we assume the same amount of power consumption by AI servers purchased from each year in each data center. Now, we vary this baseline pre-GSR configuration. We first consider the scenario where the newer-generation AI servers are dominant. More specifically, before GSR, the ratio of power consumption by the AI servers in the three years of (2021, 2022, 2023) is 1:2:3. As illustrated in Figure 8, we can still observe a large savings by GSR in terms of the carbon emission, water consumption, and electricity cost. Compared to the uniform setting in our main experiment in Section 3, the savings achieved by GSR slightly decrease. The reason is that there are relatively fewer old (or energy-inefficient) AI servers.

Next, we consider the scenario where the older-generation AI servers are dominant. More specifically, before GSR, the ratio of power consumption by the AI servers in each data center in the three years of (2021, 2022, and 2023) is 3:2:1. We show the results in Figure 9 and see that GSR can still offer substantial savings in terms of the carbon emission, water consumption, and electricity cost.

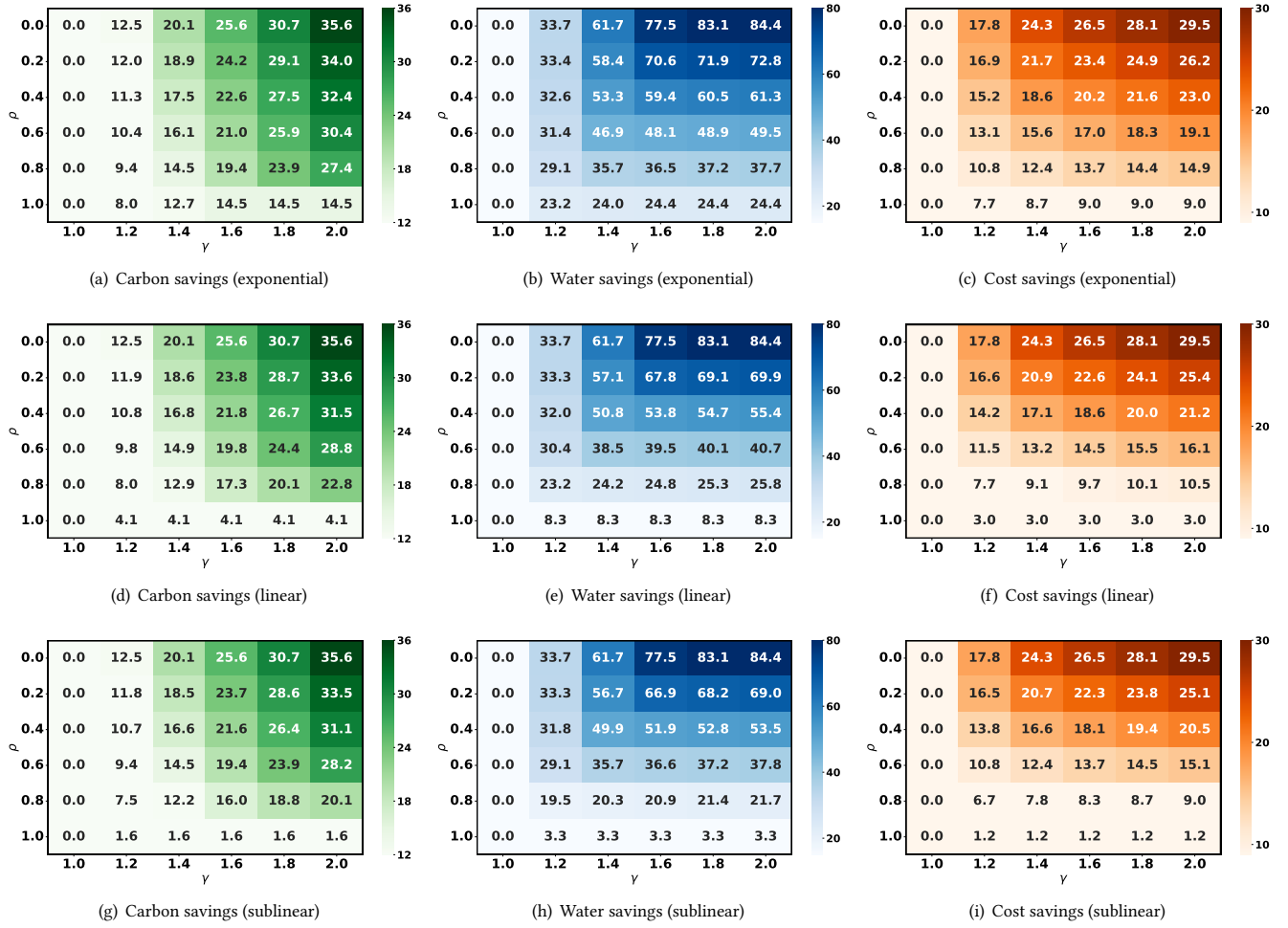


Figure 7: Carbon emission, water consumption, and electricity cost savings under different γ and ρ assessed by the synthetic GPU performance-to-power data over the last 3 years (2021, 2022, 2023).

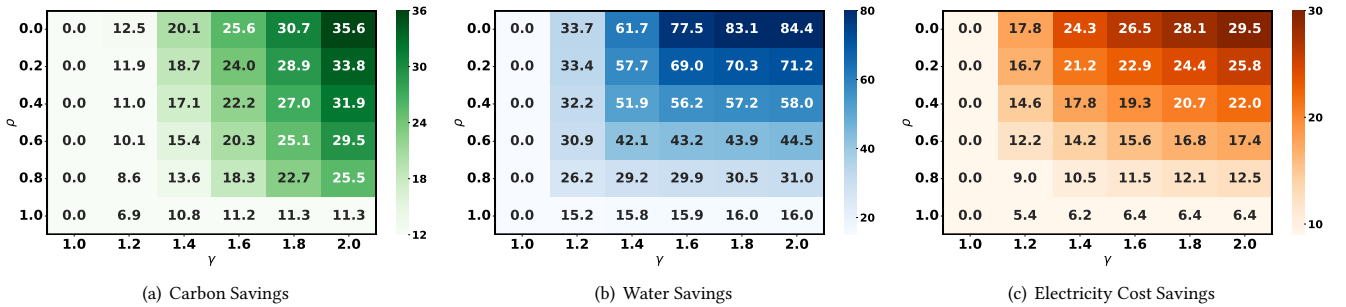


Figure 8: Carbon emission, water consumption, and electricity cost savings under different γ and ρ assessed by the manufacturer-reported GPU performance-to-power data over the last 3 years (2021, 2022, 2023). The pre-GSR ratio of power consumption by the AI servers in these three years is 1:2:3.

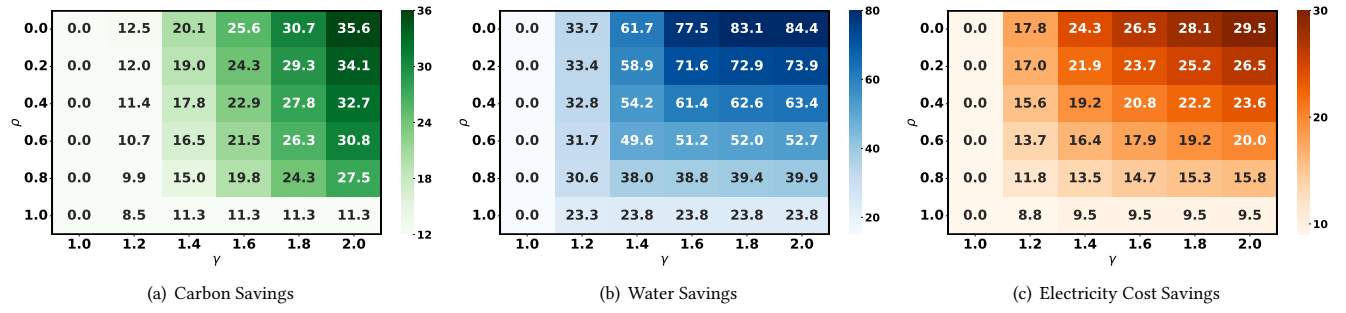


Figure 9: Carbon emission, water consumption, and electricity cost savings under different γ and ρ assessed by the manufacturer-reported GPU performance-to-power data over the last 3 years (2021, 2022, 2023). The pre-GSR ratio of power consumption by the AI servers in these three years is 3:2:1.