Minimizing Carbon Emissions by Improving Water and Energy Use Efficiencies in AI Servers: A Green Cloud Computing Strategy for Sustainable Artificial Intelligence Systems

Sumit Saklani
Department of Computer Science & Engineering
Graphic Era Hill University, Dehradun

Devendra Singh
Department of Computer Application
Graphic Era (Deemed to be University), Dehradun

Abstract:- The advent of Artificial Intelligence systems, in particular of generative models like ChatGPT, has resulted in one more area requiring heavy computational resources which in turn consumes a lot of energy and water. By estimations, one interaction with ChatGPT for instance will take an estimate of 2.9 watt hour, which is ten times higher than the amount of energy consumed to conduct an ordinary googling task that is 0.3 watt hours. This stands as a call for action toward improving the water to energy ratio of the AI systems and therefore the recent carbon emissions. This paper explores the energy efficiency patterns of AI languages such as chatbots compared with the other means of searching the internet like Google and how the effects of the AI machines on the environment can be reduced. In this connection, green cloud computing methods have been suggested as possible solutions that can be effectively combined with the principles of clean energy use; on this list are both advanced systems for maintaining low temperatures and the optimization of AI systems. Finally, using of resources could also play a crucial part in the ultimate decrease in the adverse effects that the AI industry has on our environment.

Keywords:- *AI Servers, Water Usage, Energy Efficiency, Carbon Emissions, Green Cloud Computing, ChatGPT, Google Search.*

I. INTRODUCTION

The emergence of AI models like Chatgpt has transformed human-interactions and expanded the scope of natural language processing applications. Unfortunately, these models present some impracticalities including but not limited to addressing the environmental footprint of AI models kinetic energy consumption and water use. The processes included in AI modelling incur heavy capacity hence high energy costs. A single query to the ChatGPT AI amasses an approximate energy expenditure of 2.9 watt-hours, about twenty such energy expenditure of 0.3 watt-hour as a google search.

This high energy use raises the operational expenditure and increases the greenhouse gas emissions thereby making it vital to move their embrace better operational sustainability with immediate effect. [1]. The concept of cloud computing aims at passing the benefits of AI based systems manifested in heat generation which at the cloud is further leant but concentrates on the reduction of energy as a whole through technological advances from new cooling technologies to using clean energies. The objectives of this research are the energy and water usage estimates for web-related functions of AI models such as ChatGPT and traditional search engines such as Google, and ways to improve AI server efficiency for the sake of being green.

II. LITERATURE REVIEW

A. Energy Consumption in AI Models vs. Search Engines

Models like GPT-3 and GPT-4 (Generative AI), have a sophisticated architecture resulting in an even more immense number of parameters required for text processing and synthesis, making such models considerably energy-intensive. Further, according to the research, energy demands during AI training and inference phases vastly exceed those associated with conventional computational activities, like search engine queries. [2]. Running a single inference query on ChatGPT uses 2.9 watt-hours, about ten times the energy of a Google search query (thought to be about 0.3 watt-hours). [1]

B. Environmental Impact of Data Centers

In practice, that means powering the AI infrastructure in data centers and worldwide this adds up to about 1% of all electricity consumption. [3]. The large amount of energy that data centres use comes from computing tasks and the need for cooling equipment to keep things from getting too hot. The main ingredient in traditional cooling methods is water, and huge data centres use huge amounts of it every year. [4] On the other hand, search engines like Google have switched to more energy-efficient ways to run their data centres, which has a positive effect on the environment generally.

C. Green Cloud Computing for AI Systems

Green cloud computing is now an important way to reduce the damage that cloud services do to the world. Using hardware that uses less energy, allocating resources dynamically, and getting power from green sources are all important ways to lower AI models' carbon footprint. [5]. Using air and liquid immersion cooling for AI servers has proven efficient at lowering water usage and greenhouse gas emissions, due to very high energy demands. [6].

D. Carbon Emissions from AI and Search Engines

The model training phase in AI models release significant carbon emissions A study by Strubell et al. A study by Strubell et al. (Strubell et al., 2019) estimated that training one major AI model could emit as much CO2 as five cars do over their entire lifetime. [7]. While AI models like ChatGPT are vital for tasks that require sophisticated understanding of natural language, they generate a much higher carbon footprint per search than resource-efficient search engines designed for basic queries.

III. CURRENT SCENARIO

This rapid progress for models like ChatGPT, which leverage AI have strained the need for large data center to support heavy computational workloads. AI models, particularly in the training phase, tend to do a lot of carbon emission. Strubell et al.-led researchers According to Strubell et al. (2019), training a single large AI model could result in the emissions correspondings those of five lifetime cars running. ChatGPT ate 2.9 watt-hours from recent estimates, whereas one google search query fetches as less as 0.3 watt-hours. [1]. Most of the energy consumption variability is due to the complexity of language models being powered by AI, meaning that requests and search responses have to be processed and generated with more resources than usual.

Moreover, statics of data center cooling efficiency which is important for the maintenance of necessary conditions for the operation of equipment is mainly achieved by means of water- based methods. Water over cooling systems are also installed in the data centres and these systems use millions of gallons of water every year thus making areas that are already water- scarce more deficient. [4].

In contrast to this trend, Google, along with other search engines, has significantly decreased the energy and water used per search by introducing new, more efficient cooling technologies and systems for energy management. [8].

IV. PROPOSED SOLUTION FOR MINIMIZATION WATER AND ENERGY USAGE

A. Energy Efficiency

- *AI Model Optimisation*

Optimising AI models implies for instance tweaking the architecture of models like GPT to make them less computationally hungry with same or similar performance. In that case, instead of creating more efficient approximate models through abstraction and neural architecture transformations, the techniques are now limited to using model distillation, pruning or quantization methods for energy-efficient training as well as inference. [9]. This technique greatly reduces the energy consumption per query in AI models, thus making ChatGPT not that far behind conventional search engines.

- *Dynamic Resource Allocation*

AI servers can be enhanced with dynamic resource allocation, in which computing resources are assigned on the basis of current requirements. Employing these strategies ensures that power wastage is limited throughout low demand periods as servers only run at optimal capacity if and when the situation calls for it. [10]. Google have implemented dynamic resource allocation in its data centres achieving lower energy consumption per search request. [11].

- *Sustainable Energy Integration*

It is vital to utilize energy that is derived from sustainable sources in order to minimize the carbon footprint of AI systems. Adopting this approach of sourcing from wind, solar, or hydroelectric powered data centres allows them to reduce their greenhouse gas emissions significantly. A lot of cloud service providers, like Google Cloud and Microsoft Azure, have started to use renewable energy in their processes. [12].

B. Water Usage Reduction

- *Air Cooling Technologies*

Right now, AI computers use water-based cooling systems, which don't work well and can't last for a long time. Using air cooling systems is one way to cut down on water use by a large amount. Google has created new air-based cooling solutions for some data centres, which means they don't need as much water and can save money on costs. [13].

- *Liquid Immersion Cooling*

It accomplishes this through liquid immersion cooling than submerge the server components in a electrically nonconducting, chemically inert and thermally conductive liquid, which makes for great energy and water savings. The approach is much more effective at cooling, since heat is distributed across an entire surface area in lieu of water-driven cooling. [6].

- *Water Recycling*

Closed-loop water recycling systems can be employed in data centres to reduce overall water usage. These systems continuously recycle water used for cooling, hence reducing waste and decreasing dependence on external water sources [14].

V. COMPARATIVE ANALYSIS: ChatGPT vs. GOOGLE SEARCH ENGINE

The following table shows the energy utilization and water used per query for ChatGPT and Google Search:

Service	Energy Consumption (WattHours/Query)	Water Usage
ChatGPT	2.9 W-h	High
Google Search	0.3 W-h	Low

The chart below shows that ChatGPT uses nearly 10x the amount of energy as a single Google search query. The energy used is high as this comes with the model complexity along with a need to process significant amounts of data. In addition, AI data centers are also used at a place which has very high water usage so to reduce its water footprints, whereas Google is using air cooling and recycling for water in its AI data center. [13].

VI. FUTURE EFFECTS AND SUSTAINABILITY

Green cloud computing practices could be employed to a great extent to reduce environmental repercussions from AI server hosts. It can lead to conservation of water and carbon emissions through its improved efficiency in water and energy in AI services making it environmentally friendly. Data centres can use cooling systems that are less energy-intensive and the environments must move to renewable energies or a better use of this.

Advances in optimisation techniques can be expected to close the energy gap between ChatGPT and traditional search engines. By investing in green computing R&D moving forward, the AI industry can help to combat climate change and meet worldwide sustainability goals.

VII. CONCLUSION

The research underscores the resource consumption disparity between conventional search engines (e.g., Google) and AI models like ChatGPT Search engines seem to perform better just at the level of natural language parsing, but artificial intelligence models have a far larger reach across the web. Deployment of green cloud computing via approaches like improving AI models, connecting with renewable energy sources and adopting cutting-edge cooling technologies can cut down emissions from such AI computers. Using such solutions, the AI business can effectively reduce its environmental footprint and unlock a greener future.

REFERENCES

[1]. OpenAI. "AI Model Efficiency Report," 2023.
[2]. A. Qureshi, R. Weber, H. Balakrishnan, and J. Guttag, "Cutting the electric bill for internet-scale systems," in *Proc. SIGCOMM*, 2009, pp. 123-134.
[3]. E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984-986, Feb. 2020.
[4]. Y. Zhang, X. Zong, C. Li, and G. Zhang, "Water consumption and conservation in large-scale data centers: A case study," *Energy*, vol. 220, p. 119652, Jan. 2021.
[5]. J. Baliga, R. W. Ayre, K. Hinton, and R. S. Tucker, "Green cloud computing: Balancing energy in processing, storage, and transport," *Proc. IEEE*, vol. 99, no. 1, pp. 149-167, Jan. 2011.
[6]. C. Stewart, R. Mena, and M. Garcia, "Impact of liquid immersion cooling on data center sustainability," *J. Cleaner Prod.*, vol. 221, pp. 304-311, May 2019.
[7]. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annual Meeting of the Assoc. for Comput. Linguistics*, 2019, pp. 3645-3650.
[8]. Greenpeace International, "Clicking Clean: Who is Winning the Race to Build a Green Internet?," *Greenpeace*, 2017.
[9]. M. Patel, P. Buch, and S. Parikh, "AI model optimization for energy efficiency," in *IEEE Int. Conf. on Green Tech.*, 2020.
[10]. A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, and E. Masanet, "United States data center energy usage report," *Lawrence Berkeley National Laboratory*, 2018.
[11]. Google, "Data Center Efficiency Best Practices," 2021.
[12]. "Microsoft Sustainability: Carbon Neutral Cloud," Microsoft, 2022.
[13]. T. Miller and J. Griffin, "Cooling technologies and water conservation in data centers," *J. Water Resour. Manage.*, vol. 32, no. 4, pp. 905-917, Apr. 2018.
[14]. D. Cooley and P. Gleick, "Water Recycling in Data Centers: Sustainable Practices," *Pacific Institute*, 2020.