Energy Efficient Could-Computing

1. Introduction

Cloud computing has become an integral part of modern information technology infrastructure, offering scalable resources and services over the internet. From traditional search engines to the newly developed AI models like chatGPT, the exponential growth of cloud services has led to the establishment of large-scale data centers worldwide. These data centers consume substantial amounts of energy, leading to environmental concerns and increased operational costs. This project explores the energy consumption of big data centers, the necessity for energy-efficient computing, the potential of transferring computational loads from CPUs/GPUs to dedicated hardware like FPGAs and ASICs, the challenges associated with this approach, along with some other methods to improve energy efficiency in cloud computing.

Their energy consumption has been a subject of growing concern. The Electric Power Research Institute (EPRI) points out that data center deployment, which is now greatly driven by the new AI services, has becoming a significant part of the electricity demand growth. The EPRI also estimates that data centers could consume up to 9% of U.S. electricity generation annually (about 400 TerraWh) by 2030.[1]

The market trend also shows the sharp increase in data center demand. Starting from 2022 AI services has become the major driving force for data center demand. As shown in Figure 1, the nominal price for data center rentals increases by 27%.[2] After adjusting for inflation, the real price still increases by 19%. [3]

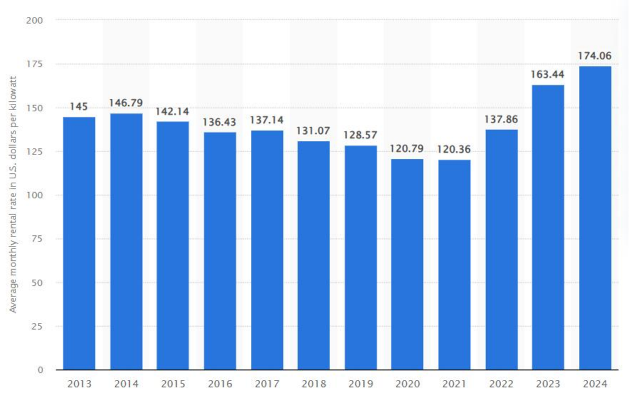


Figure 1: Avg. Monthly rental of data centers in dollars/kW

The surge in rental rates reflects the increasing pressure on data center capacity due to the rising demand, particularly emerging AI services. In recent years, high energy consumption and environmental pollution in data centers have become an urgent issue.

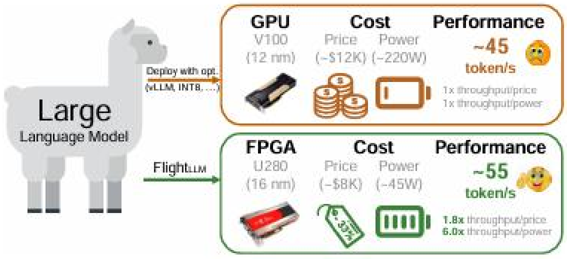
2. Hardware Solution

2.1 FPGA Implementation

A more advanced FPGA chip can take more complicated tasks than just the preparation of data. There are many FPGA implementations of the transformer model, which is the key architecture of many modern NLP and CV models. For example, H.Rong, et al proposed TransFRU, which is a FPGA-based accelerator for the self -attention mechanism. It achieved a 5.9x energy efficiency compared with GPU [4].

Microsoft has started several projects (e.g. Project Brainwave) that aimed to implement Neural Networks on FPGA since 2015.[5]

FlightLLM, an efficient LLM inference implemented on FPGA, shows a strong performance and energy efficiency.[6]



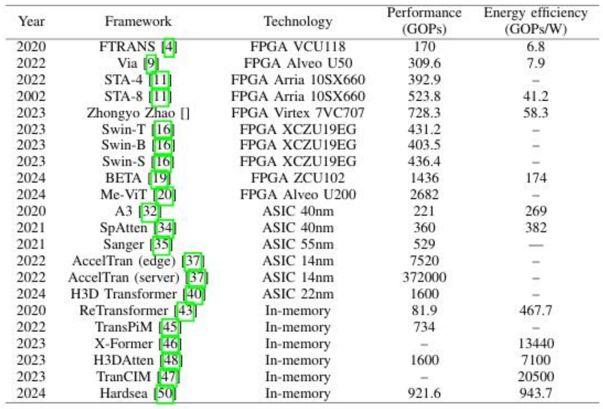
From the Figure the U280 FPGA outperformed V100 GPU in cost (both fiance and energy) and performance.

However this direction is facing challenges as current AI models are getting bigger in size which is exceeding the capabilities of FPGA. V100 is a GPU released in 2017, today's GPU like A100 (2-3 times faster than V100) are much powerful than V100, which leaves only the power advantage to the FPGA solution. The development of FPGA technology is slower than GPU development. Furthermore, industry prefer stacking with GPUs as it is a simpler and more comfortable way, given the energy cost is not a very big concern compared to other costs.

2.3. Other hardware solutions

Besides FPGA, there are many other hardware techniques that aim to address this issue, such as ASIC and In memory computing. Koilia and Kachris performed a detailed survey on the performance of LLM on various hardware. [7]

TABLE I
LLM-TRANSFORMER ACCELERATORS



Year	Framework	Technology	Performance (GOPs)	Energy efficiency (GOPs/W)
2020	FTRANS [4]	FPGA VCU118	170	6.8
2022	Via [9]	FPGA Alveo U50	309.6	7.9
2022	STA-4 [11]	FPGA Arria 10SX660	392.9	–
2002	STA-8 [11]	FPGA Arria 10SX660	523.8	41.2
2023	Zhongyo Zhao []	FPGA Virtex 7VC707	728.3	58.3
2023	Swin-T [16]	FPGA XCZU19EG	431.2	–
2023	Swin-B [16]	FPGA XCZU19EG	403.5	–
2023	Swin-S [16]	FPGA XCZU19EG	436.4	–
2024	BETA [19]	FPGA ZCU102	1436	174
2024	Me-ViT [20]	FPGA Alveo U200	2682	–
2020	A3 [32]	ASIC 40nm	221	269
2021	SpAtten [34]	ASIC 40nm	360	382
2021	Sanger [35]	ASIC 55nm	529	—
2022	AccelTran (edge) [37]	ASIC 14nm	7520	–
2022	AccelTran (server) [37]	ASIC 14nm	372000	–
2024	H3D Transformer [40]	ASIC 22nm	1600	–
2020	ReTransformer [43]	In-memory	81.9	467.7
2022	TransPiM [45]	In-memory	734	–
2023	X-Former [46]	In-memory	–	13440
2023	H3DAtten [48]	In-memory	1600	7100
2023	TranCIM [47]	In-memory	–	20500
2024	Hardsea [50]	In-memory	921.6	943.7

The survey points out that In-memory computing is extraordinary in terms of Energy efficiency. However, given the current high cost and its immaturity compared to the other 2 methods, memory-computing is not as widely used as FPGA or ASIC. ASIC solution achieves the best performance. But the cost can be major concern, especially given that limited demand for these chips. However, the rapid growth in data centers is mitigating this issue and ASIC is gaining a larger portion in industry. FPGA stands as a middle choice, yet its simplicity and relatively lower cost makes it the most common choice.

Although GPU is still the major choice in the current industry landscape, FPGA and ASIC are being actively employed to accelerate large AI models, each offering unique benefits. In-memory computing holds promise for future AI acceleration but is still in the developmental stage and not widely adopted in the industry.

3. Non-hardware Solution

3.1 Software Optimization

Optimizing software to be more efficient can reduce the computational resources required, thereby lowering energy consumption. Software Defined Networking (SDN) paradigm has shown a great penitential in optimizing routing and flexibility in network management [8]

3.2 Edge Computing

Another direction of reducing load on cloud server is to move some of the tasks to the client side, which is accomplished through edge computing. While edge computing has proven its ability to effectively reduce network traffic and ease the burden on cloud data centers by handling local computations, efficiently scheduling tasks in a hybrid edge–cloud environment remains a significant challenge. Some research teams are exploring more efficient scheduling algorithms. For example, K.Zhu's team proposed a graph neural network based networks combined with reinforcement learning techniques that can significantly improve the quality of service[9]

4. Conclusion

Energy-efficient cloud computing is essential for minimizing the environmental impact of data centers and reducing the operational costs associated with energy consumption. The current energy usage of data centers is unsustainable, and the need for energy-efficient solutions is more pressing than ever. Transferring computational loads to dedicated hardware such as FPGAs and ASICs offers significant energy-saving potential, but also poses several challenges, including high development costs, complexity in integration, and competition from powerful CPU/GPUs.

Other approaches, such as optimizing software and networking, using renewable energy sources, and developing edge computing, have also been explored to achieve energy efficiency in cloud computing. Future research should focus on overcoming the challenges associated with dedicated hardware and further exploring the potential of combining multiple approaches to develop a holistic solution for energy-efficient cloud computing.

Reference

[1] U.S. Department of Energy, "Clean Energy Resources to Meet Data Center Electricity Demand,"Energy.gov,[Online].Available:https://www.energy.gov/policy/articles/clean-energy-resources-meet-data-center-electricity-demand.

[2] “U.S. Data center rental rate 2024 | Statista,” Statista, 2024. https://www.statista.com/statistics/1370191/data-center-rental-rate-us/ (accessed Nov. 2024).

[3] Bureau of Labor Statistics, “CPI Inflation Calculator,” Bls.gov, 2024. https://data.bls.gov/cgi-bin/cpicalc.pl

[4] H. Wang, Y. Bai, J. Yu, and K. Wang, “TransFRU: Efficient Deployment of Transformers on FPGA with Full Resource Utilization,” Jan. 2024, doi: https://doi.org/10.1109/asp-dac58780.2024.10473976.

[5] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. Chung, “Accelerating Deep Convolutional Neural Networks Using Specialized Hardware,” 2015. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/CNN20Whitepaper.pdf

[6] S. Zeng et al., “FlightLLM: Efficient Large Language Model Inference with a Complete Mapping Flow on FPGAs,” arXiv.org, 2024. https://arxiv.org/abs/2401.03868 (accessed Nov, 2024).

[7] N. Koilia and C. Kachris, “Hardware Acceleration of LLMs: A comprehensive survey and comparison,” arXiv.org, 2024. https://arxiv.org/abs/2409.03384 (accessed Nov, 2024)

[8] Beakal Gizachew Assefa and Oznur Ozkasap, “A survey of energy efficiency in SDN: Software-based methods and optimization models,” vol. 137, pp. 127–143, Jul. 2019, doi: https://doi.org/10.1016/j.jnca.2019.04.001

[9] K. Zhu, Z. Zhang, Sherali Zeadally, and F. Sun, “Learning to Optimize Workflow Scheduling for an Edge - Cloud Computing Environment,” IEEE Transactions on C loud Computing, vol. 12, no. 3, pp. 897-912, May 2024, doi: https://doi.org/10.1109/tcc.2024.3408006.