Title: The Potential of Data Center Energy Demand to Provide Grid Flexibility

Introduction

The growth of AI has led to a rapid increase in energy demand from data centers [1]. Current estimates suggest that data centers will consume about ~10% of the nation's electricity demand by the end of this decade [2]. The modern electric grid has not previously seen such sudden increases in electricity demand, and data center owners are increasingly concerned about the ability to power their data centers [3].

Integrating such data centers in an electricity network powered increasingly by variable renewable energy provides both additional challenges and opportunities. Many data centers have inherent flexibility with how and where they process jobs (and hence consume power), and as such there has been recent interest in using this flexibility to provide balancing services to the electric grid.

In this paper we describe the current state of the research in this area and assess the evidence that data centers can provide grid flexibility services.

Background on Data Centers

Data centers house large, centralized computing, storage, and networking resources. Typically, data centers can be subdivided into internet and computing data centers. Internet data centers handle tasks necessary for everyday services such as streaming of videos (YouTube, TikTok), routing internet traffic, storage of user data (e.g. Google drive), processing database queries (e.g. Facebook), and communicating player interactions during online gaming. All these tasks require high throughput networking capability, low latency, and some storage capability. Processing capability provided by central processing units (CPUs) is typically not a significant constraint in these data centers. On the other hand, compute data centers handle compute intensive tasks such as numerical simulations, training and inference of AI models, crypto-currency mining, and archival storage of large datasets. These centers house large numbers of graphics processing units (GPUs), CPUs, and application-specific integrated circuits (ASICs) distributed over 1000s of nodes, which require cooling to operate within a certain temperature range. Furthermore, they often feature a high bandwidth low-latency interconnect to allow rapid exchange of data between these nodes.

Data centers are typically owned by private companies such as Google or government entities such as the Department of Energy (DOE) or National Science Foundation (NSF). They can house hardware that is dedicated to a single customer or shared among customers who pay credits for utilizing shared resources. These resources may be assigned to virtual machines accessible to users, or users may submit jobs to a shared queue. While virtual machines provide many benefits including enhanced customization and safety to consumers, they introduce overhead and are thus rarely utilized on government owned supercomputers dedicated to academic research. To simplify this article, we won’t discuss virtual machines further and assume that resource requests (user-submitted jobs) are handled by a queue manager (scheduler). Many data centers have specialized queues for debugging, which give priority access for short periods of time at a higher cost, and queues that provide discounts. The discounts are given at the cost of longer wait times or the ability of the data center to prematurely take back the allocated resources.

A scheduler subsequently then assigns resources to the submitted jobs. While many government owned supercomputers must allocate individual jobs to entire nodes (each node comprising many processing cores), many commercial data centers offer more flexibility and allow users to request access to individual CPU cores and/or GPUs. A node in modern data centers typically consists of ~100 CPU cores and/or between 1-8 GPUs. Each node usually also has a small amount of fast local storage (~several Terabytes) and can access a centralized file system (~hundreds Petabytes of data). Some data centers are homogeneous, while others offer many different nodes featuring different CPU/GPU architectures and vendors.

In the past, compute data centers were primarily tasked with running large simulation and data analysis code. Today, their use is now dominated by AI tasks. These tasks can involve training of machine learning models or usage of (inferencing) a pre-trained model. Training tasks are significantly more expensive, often require 1000s of GPUs, and can take days or weeks to run. On the other hand, inferencing a pre-trained model (e.g. using ChatGPT) uses a relatively small amount of compute power with much fewer GPUs, but can be run by millions, and in the future, billions of users. Recent estimates suggest that with the popularity of large language models, the total energy consumption of inference now exceeds that of training (given the large number of users)[4].

We expect that the power consumption attributable to datacenters to increase over the coming years, driven by an increase in the number of data centers [5], and by GPUs consuming more power. For example, NVIDIAs V100 GPU released in 2017 had a thermal design power (TDP) of 300W while NVIDIA’s next generation B200 GPUs to be released in 2025 will have a TDP of 1200W. This trend is expected to continue as advances in semiconductor manufacturing will allow hardware manufacturers to cram an exponentially increasing number of transistors into individual CPUs/GPUs. At the same time, limits in physics will make it challenging for hardware manufacturers to decrease the power consumption of transistors proportional to the increase in transistor density.

Background on Grid Flexibility

Over the past decade there has been significant effort worldwide to decarbonize the power sector, largely through the deployment of variable renewable energy. Solar and wind power represent cost competitive alternatives to conventional carbon-based generation, but their outputs are inherently dependent on weather (variable and uncertain). Coupled with the rapid increase in power demand (from AI as well as transport and heating electrification) more widespread adoption of renewables is leading to a volatility in the supply and price of electricity [6].

Load flexibility is a potential avenue to offset the variability of renewables’ output, where the timing and/or location of power demand is altered to better align with available generation. A recent study suggested that if system operators allowed for a relatively small number of interruptions, up to 76GW of new load could be added to the US system without increasing generation [7]. Given the first order analysis conducted, this figure should be treated as an upper bound.

Power system operations can utilize load flexibility in several manners, two distinct modes are:

(1) **Regulation/Reserve Services.** Devices providing this service are paid a retainer fee, in return for a guarantee that power will be provided within a set time period (typically seconds to minutes) of receiving a signal. These signals are rare, however the services are contracted in advance and there are large penalties for failing to respond. These services are designed to avoid power loss during extreme events caused by component failures.

(2) **Energy Arbitrage.** While regulation/reserve services focus on rare quick response events, energy arbitrage focuses on routine moving energy over coarser time and/or spatial scales For example, grid scale batteries may perform arbitrage by charging when surplus energy is available and discharging during peak times. These services tend to be less profitable, but available frequently, and do not require contracts in advance. Arbitrage provides benefits to the grid (and customers) by allowing system operators to solve routine congestion problems or better utilize low-cost generation.

Data Center Flexibility Mechanisms

Various control regimes have been proposed to introduce demand flexibility into the operation of data centers. In this article we primarily focus on compute data centers since these show the most rapid growth in total power consumption, but our conclusions should be generally applicable.

Uninterruptible Power Supply

One proposed flexibility mechanism involves the use of the data centers' protection system for regulation or reserve services [8,9]. Uninterruptible power supply (UPS) protection systems are designed to keep the data center online for a short period (hours) in the event of power outages, and as such have extremely low utilization. However, to provide regulation services the center must commit to prioritize the service over local power supply in an emergency event.

Rerouting jobs

Most of the research into the flexibility potential of data centers focuses on transferring requests between data centers. This would move power demand to other locations within the power grid, and potentially allow one center to completely shut down, reducing total energy consumption [10,11]. Early work focused on internet data centers, but later work has extended the idea to HPC data centers, including those focused on AI. HPC data centers typically involve management of queued jobs and thus have the potential to delay processing of a job. This allows for the movement of the jobs in time as well as space (increasing the flexibility potential); however, these data centers also operate at higher levels of utilization given that they can queue jobs [12,13,14].

Smart scheduling

Compute data centers often maintain queues for jobs; therefore, it is likely that some degree of arbitrage should be possible via better management of the queue. Current job schedulers prioritize the timely and efficient execution of the submitted jobs in a queue. Often such schedulers optimize for achieving a high utilization rate of the data center. Some recent work also highlights the benefit of using "smart" schedulers to move the most demanding jobs to off peak hours [15,16]. However, all this work has been so far focused on the jobs that are currently submitted to data centers and does not exploit the potential flexibility users could offer to the datacenter.

User-provided flexibility in the job size, duration and hardware configuration used can potentially dramatically increase the utilization of data centers and make them amenable to demand-response strategies by moving power-hungry jobs to off-peak hours. For example, many HPC-based simulation codes and AI workloads have the capability to utilize various types of hardware and job sizes. Furthermore, they checkpoint regularly, so can be terminated quickly if the need arises. On the other hand, jobs tasked with AI inferencing need to be able to process user queries on very short timescales, leading to a low average utilization rate.

Clock rate modulation

There are other, less studied mechanisms by which a data center could modulate its power consumption. Firstly, the clock rate and voltage of the CPU and GPU cores can be reduced to effectively slow down the processing speed of GPUs, lowering its power consumption. This works because, for most codes, the reduction in computational performance is much smaller than the reduction in power consumption. To test the feasibility of reducing the clock speed, we have tested underclocking a single A100 GPU running the H-AMR general relativistic magnetohydrodynamics (GRMHD) code [17] used in astrophysics and we achieved a 40% drop in power consumption at a 22% drop in raw performance.

This shows that underclocking provides an avenue to reduce the power consumption of data centers that are underutilized. For example, jobs can be executed at a lower clock frequency, but run for longer, achieving more computations per unit energy, and still getting all the work done. However, this possibility is not typically exploited, since users submit fixed-duration jobs and expect to get consistent performance. Besides increasing the efficiency of underutilized data-centers, clock speed modulation could be used as a wild-card available to data center operators to significantly reduce their power consumption during extreme events [18].

Pre-cooling

Data centers expend 20-40 percent of their energy on cooling. While air cooling is the norm in internet data centers, liquid cooling (where water flows through metal radiators extracting heat from the hardware), and immersion cooling (where the hardware is immersed in a specialized liquid) is becoming the norm in datacenters housing a high density of GPUs. For example, NVIDIA's B200 GPUs have a TDP of 1200W and a single rack will be able to fit 72 of such GPUs leading to a total power consumption of 120 kW per rack. Several data centers already reduce power consumption during peak hours by pre-cooling a large volume of coolant [1, 19].

Flexible Job Characteristics

Jobs executed in a single data center can show a large variability in power consumption. This depends on several factors including how efficiently the algorithm utilizes the compute hardware and its bottlenecks, as well as the hardware itself. For example, a highly optimized simulation code explicitly targeting a specific hardware architecture will be able to utilize the hardware more efficiently than a less-optimized community code that relies on compiler directives to target specific computer architectures. Similarly, executing a job on newer hardware using smaller transistors typically uses less energy than executing the same job on older hardware with bigger transistors.

Power consumption of computing jobs is also determined by the type of instructions they execute. These instructions can be broken down in tensor core operations that multiply matrices, floating point operations, integer operations, and memory read/write operations. Tensor core operations utilized by AI workloads are the most power hungry, followed by floating point operations utilized by many non-AI HPC workloads. This implies that jobs that are limited by memory or networking bandwidth will use less power than compute intensive jobs. Furthermore, large jobs utilizing a large number of nodes tend to be less efficient than smaller jobs due to communication overhead and load imbalance. For example, training a large language model on more GPUs than required due to memory constraints can increase the training cost several fold [20].

Cryptocurrency mining

Recently, there has been significant interest in cryptocurrency mining as a source of flexible energy demand [21-24]. Because mining workloads are predictable, homogenous, and relatively decoupled from external dependencies, they offer a more tractable problem for optimization and scheduling models. However, at current electricity prices and cryptocurrency values, GPU-based mining is largely unprofitable. Instead, most operations now use highly specialized hardware such as ASICs, which offer much higher efficiency per watt. Cryptocurrency mining facilities are therefore optimized to house the maximum number of ASICs at minimal cost, and unlike traditional data centers, they typically do not include high-end networking, storage, or resilience features.

Quantifying flexibility potential

For each of the discussed flexibility mechanisms, the ability of a data center to participate depends on the specific hardware and job constraints. For example, to provide spatial arbitrage (reduce demand in a location of the grid experiencing scarcity while increasing demand at a different location), a company needs to have multiple datacenters with the ability to handle the same request at the same time. Regarding temporal arbitrage, relatively little research has investigated the potential scheduling flexibility of current workloads and codes running in data centers. While some data centers are shifting loads to low demand times, this usually only involves shifting non-urgent fixed duration jobs in time and does not include changing the job characteristics [25].

The ability of a data center to provide flexibility will depend on a variety of factors including the utilization rate of the data center, the variability in power consumption of different jobs, and the flexibility these jobs offer to the smart scheduler. A recent paper attempts to quantify the flexibility of 14 HPC prices [26]. They use computing traces to estimate flexibility and suggest that AI-focused data centers may be especially well suited to provide flexibility (compared to general-purpose clusters). These findings are based on optimization models to approximate the re-scheduling possible with maximum delays of 10 or 20%. They focus on defining average flexibility cost, assuming that any of the jobs could theoretically be delayed (thus can be seen as an upper bound estimate).

Another study surveyed data center owners to attempt to understand the barriers to data centers providing flexibility in China [27]. In general, data center owners were resistant to the idea; one respondent expressed that very few data centers are sensitive to electricity price compared to revenue from processing jobs; another stressed that the data centers paid a lot for their UPS systems and would therefore be unwilling to deploy it for ancillary services as their power supply may then be threatened in an emergency.

Incentivizing flexibility

In order to deploy flexibility in data centers, there need to be sufficient incentives for both users to participate and data centers to utilize flexibility for grid resilience.

Incentivizing users to provide constraints

In order to deploy flexibility without consumer dissatisfaction, schedulers need information about the requirements and urgency of users' jobs. While current job schedulers provide some flexibility for the user to select variable job sizes, runtime durations, and compute hardware, data centers do not incentivize the user to make use of such smart scheduling features. Usually the only provided incentive to the user is a vague promise that the job will run sooner if it is able to backfill into an existing queue. Hence, very few users make use of these scheduling features.

In order to realize additional flexibility, we propose that providing dynamic prices to users has the potential to increase the efficiency of data centers. In this paradigm, the user submits jobs with certain characteristics to the data center and specifies how much they are willing to pay if these jobs are launched on different hardware or on a different number of nodes than what they originally requested. On the other hand, data centers can specify dynamically how much of a discount they are willing to provide for a job to run with non-ideal characteristics. This will lead to a market-like dynamic where users are encouraged to request reasonable discounts and data centers are encouraged to honor such requests.

For example, with current job scheduling policies users might be resistant to run on more nodes or less efficient hardware because a slightly shorter queue time (the only incentive offered on many clusters) does not weigh up against the increased cost to run their codes. Hence, these users currently do not provide this flexibility to the scheduler. However, if such users can specify what fraction of the normal cost they are willing to pay when accepting non-ideal job characteristics, a data center can capitalize on this added flexibility provided by the user. Smart schedulers that can capitalize on user and data-center provided constraints in this manner are not developed yet, and they will require involvement of users and datacenters to be successfully integrated in their workflows.

Incentivizing data centers to provide grid flexibility

Implementing better scheduling algorithms to provide flexibility does not guarantee that data centers will use this flexibility to shift their power consumption. The majority of papers on the topic of data center management focus on the use of electricity price to motivate energy arbitrage. However, researchers have found data center operators to be fairly insensitive to electricity price [27]. Most work which quantifies the "price" at which data centers would provide the flexibility considers only the direct costs to the center operator [26]. However, from the perspective of the center operator, the benefit from providing flexibility needs to outweigh the opportunity cost of taking on more jobs. The superior scheduling algorithms discussed would also allow the center to better utilize its own resources, and therefore possibly increase the total power consumption.

On average a GPU is rented out at $2/hr (with potential discounts for large users), while average usage figures from OLCF Frontier implies each GPU consumes an average of only 0.32 kW of power [28]. Using average energy prices, we can see that electricity should account for less than $0.04 compared to the $2 revenue. Furthermore, the electricity price at which the energy costs outweigh the job profits is $6,250/MW – a price which has only historically been hit during extreme events such as the Texas blackout. This suggests that using only electricity prices as incentive data centers would almost always be incentivized to take additional jobs rather than provide flexibility. If data centers implement superior scheduling algorithms, they are likely to use them to more efficiently (heavily) utilize their existing systems rather than alter their energy consumption. Bitcoin mining may be more price sensitive, using today's Bitcoin value the energy cost exceeds the revenue at $790/MW – but, as previously mentioned, Bitcoin mining is rarely profitable using GPUs nowadays.

New data centers often face extortionate connection costs which they are sensitive too – as evidenced by Amazon's recent lawsuit [3]. It is possible that flexibility could be incentivized in return for a reduced interconnection cost – e.g. a contracted annual number of demand reduction calls.

Conclusion

In this paper we discussed the features of data centers, their jobs, and their suitability to provide grid flexibility. Internet data centers typically have a lot of latency and are therefore able to reroute jobs (providing locational flexibility) at very low cost. However, these data centers need to compute near instantaneous tasks, and thus can't help with the temporal variation in renewable energy.

On the other hand, compute data centers (which are concerned with AI tasks) aim to operate at near to their full capacity, keeping jobs in queues. There are a variety of mechanisms via which these data centers could increase their energy demand flexibility including: modulating clock rate, pre-cooling, and better scheduling. It is very challenging to make estimates of the total flexibility potential of data centers because very few compute data centers share public information about utilization or energy demand. In addition, due to the large number of users utilizing a data center we expect significant challenges in gathering high-quality performance data of various workloads to assess potential job scheduling flexibility.

Furthermore, implementing better routing or scheduling of jobs may chiefly enable centers to better utilize their own resources. Given the large sunk capital costs and highly lucrative jobs, we conclude that it is unlikely that centers would be routinely incentivized to perform energy arbitrage based on electricity prices alone. Our estimates of the electricity price at which a data center would deny a job ranged from $790/MW for Bitcoin mining to $6,250/MW for GPU accelerated workflows. A more promising avenue might be to negotiate discounted interconnection costs for new data centers in return for some contracted flexibility.

Cryptocurrencies mining is an edge case that has gained a lot of interest due to its predictable nature and it being relatively sensitive to electricity price. It should be noted that today Bitcoin mining mostly takes place on specialized hardware and not in traditional data centers. These specialized machines may demonstrate some energy price-following behavior in the coming years. However, these would be new centers (adding total load) and the volatile nature of cryptocurrencies coupled with their relative lack of societal benefit makes this a controversial source of flexibility.

References

1. National Renewable Energy Laboratory (NREL). Reducing Data Center Peak Cooling Demand and Energy Costs With Underground Thermal Energy Storage. 2025.
2. Jones N. How to stop data centres from gobbling up the world's electricity. Nature. 2024;618(7962):16–7.
3. The Hill. Amazon data center energy grid [Internet]. 2024. Available from: https://thehill.com/opinion/technology/4976847-amazon-data-center-energy-grid/
4. Contrary. How Much Energy Will It Take to Power AI? [Internet] Available from: https://www.contrary.com/foundations-and-frontiers/ai-inference
5. Wong, Y. How many data centers are there and where are they being built? ABI Research. 2024.

6. Wang L, von Laszewski G, Younge AJ, He X, Kunze M, Tao J, et al. Cloud computing: a perspective study. New Gener Comput. 2010;28(2):137–46. Available from: https://ieeexplore.ieee.org/abstract/document/6197252
7. Norris T, Patiño-Echeverri D, Profeta T. Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems. Nicholas Institute for Energy, Environment & Sustainability, Duke University; 2025.
8. Li X, Zhang H, Xu X, Li Y, Zhang H, Xu X. Energy-efficient resource allocation for industrial internet of things in 5G heterogeneous networks. Appl Energy. 2021;292:116871.
9. Zhang H, Li X, Li Y, Zhang H, Li Y. A survey of 5G network: Architecture and emerging technologies. IEEE Access. 2015;3:1206–32.
10. Shang Y, Li D, Xu M. Energy-aware routing in data center network. InProceedings of the first ACM SIGCOMM workshop on Green networking 2010 Aug 30 (pp. 1-8).
11. Li X, Zhang H, Xu X, Li Y, Zhang H, Xu X. Energy-efficient resource allocation for industrial internet of things in 5G heterogeneous networks. Appl Energy. 2021;292:116871.
12. Zhou Y, Paredes A, Essayeh C, Morstyn T. AI-focused HPC Data Centers Can Provide More Power Grid Flexibility and at Lower Cost. arXiv preprint arXiv:2410.17435. 2024.
13. Majumder S, Aravena I, Xie L. An Econometric Analysis of Large Flexible Cryptocurrency-mining Consumers in Electricity Markets. arXiv preprint arXiv:2408.12014. 2024.
14. Dayarathna M, Wen Y, Fan R. Data center energy consumption modeling: a survey. IEEE Commun Surv Tutor. 2016;18(1):732–94.
15. Venkataswamy V. Job scheduling in datacenters using constraint controlled RL. arXiv preprint arXiv:2211.05338. 2022.
16. Piontek, T., Haghshenas, K. & Aiello, M. Carbon emission-aware job scheduling for Kubernetes deployments. *J Supercomput* 80, 549–569 (2024).
17. Liska, M. et al. H-AMR: a new GPU-accelerated GRMHD code for exascale computing with 3D adaptive mesh refinement and local adaptive time stepping. Astrophysics Data System. 2021.
18. Shehabi A, Smith SJ, Masanet E, Horner N. Data center growth in the United States: decoupling the demand for services from electricity use. Environ Res Lett. 2021;16(6):064028.
19. Vertiv. Evolving chilled water cooling methods for slab floor data centers [Internet]. Available from: https://www.vertiv.com/en-emea/about/news-and-insights/articles/blog-posts/evolving-chilled-water-cooling-methods-for-slab-floor-data-centers/
20. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-LM: training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053. 2019.
21. Hajiaghapour-Moghimi M, Hajipour E, Hosseini KA, Vakilian M, Lehtonen M. Cryptocurrency mining as a novel virtual energy storage system in islanded and grid-connected microgrids. International Journal of Electrical Power & Energy Systems. 2024 Jul 1;158:109915.

22. Menati A, Cai Y, El Helou R, Chao T, Xie. Optimization of Cryptocurrency Mining Demand for Ancillary Services in Electricity Markets. Proceedings of the 57th Hawaii International Conference on Systems Science. 2025
23. Carter N, Connell S, Jones B, Porter D, Rudd MA. Leveraging Bitcoin miners as flexible load resources for power system stability and efficiency. [Internet]. 2023 Nov 22.
24. Menati A, Lee K, and Xie L, Modeling and Analysis of Utilizing Cryptocurrency Mining for Demand Flexibility in Electric Energy Systems: A Synthetic Texas Grid Case Study. *IEEE Transactions on Energy Markets, Policy and Regulation. 2023; 1; 1*
25. Mehra V, Hasegawa R. Supporting power grids with demand response at Google data centers. Google Cloud Blog. 2023 Oct 3.
26. Zhou Y, Paredes A, Essayeh C, Morstyn T. AI-focused HPC Data Centers Can Provide More Power Grid Flexibility and at Lower Cost. arXiv preprint arXiv:2410.17435. 2024.
27. Li X, Zhang H, Xu X, Li Y, Zhang H, Xu X. Energy-efficient resource allocation for industrial internet of things in 5G heterogeneous networks. Appl Energy. 2021;292:116871.
28. Sun, J., Gao, Z., Grant, D. *et al.* Energy dataset of Frontier supercomputer for waste heat recovery. Scientific Data .2024; 11; 1077.