IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 June 2024, accepted 5 August 2024, date of publication 7 August 2024, date of current version 19 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3440217

RESEARCH ARTICLE

Will Energy-Hungry AI Create a Baseload Power Demand Boom?

JONAS KRISTIANSEN NØLAND, (Senior Member, IEEE),
MARTIN HJELMELAND, (Member, IEEE),
AND MAGNUS KORPÅS, (Member, IEEE)
Department of Electric Energy (IEL), Norwegian University of Science and Technology (NTNU), 7034 Trondheim, Norway
Corresponding author: Jonas Kristiansen Nøland (jonas.k.noland@ntnu.no)

ABSTRACT The rapid expansion of generative artificial intelligence (AI) technologies is projected to significantly affect electricity use in the global data center sector. Earlier research has proposed using data centers for load-balancing the future power grid to allow higher integration of variable renewables. In this paper, we review the expected future electricity consumption of AI and evaluate the behavior of AI data centers in clean energy systems. Our work found that the levelized cost of computing (LCOC) favors higher load factors and shows a relatively low sensitivity to electricity price levels. Under our baseline cost assumptions, a baseload electricity price of \$125/MWh benefits load factors higher than 64 %, depending on the market price conditions and variations. Nevertheless, high-tier data centers with higher operational costs and capital expenditures favor even higher load factors to optimize their LCOC. These findings show that a boom in AI energy use could drive significant baseload power demand in future power systems.

INDEX TERMS Artificial intelligence (AI), AI energy use, computing cost, computing efficiency, data center, graphical processing units (GPUs), load factor, load shifting.

NOMENCLATURE

α	Power usage effectiveness (PUE), [−].
γ	Computing efficiency, [PFLOPs/kW].
c	Overnight construction cost (CAPEX), [\$/kW].
d	Annual O&M cost (OPEX) of total CAPEX, [%].
k	Load factor (utilization rate), [%] or [−].
n	Number of years of depreciation, [−].
p	Average price of electricity, [\$/MWh] or [¢/kWh].
r	Weighted average cost of capital (WACC), [%].
t	Average time per year, 8765.8 h or 31 556 926 s.

I. INTRODUCTION

Artificial intelligence (AI) is inherently an energy-intensive technology, and for this reason, green AI is an emerging research field [1]. The International Energy Agency (IEA) estimates that data centers currently account for 1 to 1.5 % of global electricity demand [2]. Like other power-intensive industries, AI data centers urgently need low-carbon energy resources to meet their demands and not conflict their existence with climate goals and objectives. Consequently, integration with renewable sources has been an important strategy, and to do so, they have been oriented toward moving data centers to locations where the greenest electrons are found [3]. Nuclear energy has also recently received attention owing to its role in AI development [4].

By 2027, NVIDIA could be annually shipping 1.5 million AI server units [6], with a power demand of 6.5 to 10.2 kW per unit. Assuming a load factor of 100 %, this could result in an annual increase in AI-powered electricity consumption of 85 to 134 TWh per year [7]. Fig. 1 presents three possible IEA scenarios for the future electricity needs of the data center sector, with linear extrapolations from 2026 to 2035, extending the predictions of the IEA. Others have also predicted a base case of new yearly electricity demand of 1000 TWh for AI-related activities by 2030 [8], slightly above the IEA's base case. Nevertheless, the rising demand for AI is one of the biggest uncertainties in future energy use.

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Raza.

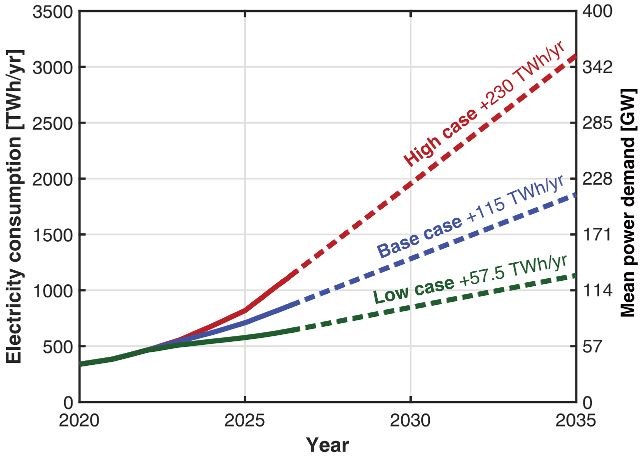

FIGURE 1. IEA's projected annual electricity consumption and mean power demand from data centers, AI, and cryptocurrencies until 2026 [5] with linear extrapolations toward 2035 for high, base, and low cases.

Masanet et al. observed that despite a massive expansion of data storage between 2010 and 2018, the data center's power demand did not significantly increase owing to efficiency gains [9]. However, this trend could change if Moore's law end, where Koot and Wijnhoven predict that an exponential power demand increase could be the result [10]. The implications are exponential extrapolations in Fig. 1 instead of linearizing future demand. However, a continued increase in the computing efficiency of AI would reduce the energy needs [11], but it could also cause a rebound effect, resulting in higher computing demand (i.e., Jevons' paradox). Such a scenario may also be likely, as there are no natural asymptotic destinations that restrict the need for the processing, storage, and transportation of data.

Another major uncertainty with AI data centers is their compatibility with load-shifting of their power demand. Krioukov et al. examined workload scheduling to make it compatible with wind power production patterns [13]. A significant increase in the wind share was obtained, but it also caused bottlenecks, with some workloads exceeding their deadlines. Similarly, Goiri et al. investigated how workload scheduling could be used to integrate more local solar energy with an electrical grid as the backup [14]. Moreover, a combination of pre-planned batches of workloads and interactive workloads was studied by Liu et al. to allow for more workload delays [15]. To go further, Chen et al. looked at the geographical distribution and load-balancing of data centers in California [16]. By routing workloads between data centers, planning can be made to minimize overall fossil fuel use. However, Lui et al. found that this approach is a trade-off between energy cost and computing performance costs [17]. Nevertheless, Toosi et al. reported that geographical smoothing can be achieved without degradation in service quality [18]. Although the potential to provide demand responses from data centers has been thoroughly discussed, Ghatikar et al. [12] pointed out that the value of data centers participating in such programs should be comprehensively assessed.

In this study, we develop a basic economic model to estimate the levelized cost of computation (LCOC), which we introduce as a new performance metric to economically evaluate the data center infrastructures used to execute AI workloads. Thus, we can evaluate the costs of running data centers part-time at low loads to balance the power system. Our fundamental research question is whether future data center infrastructure will predominantly favor baseload demand or if the infrastructure will also provide economic opportunities for flexible demand and load-shifting, reducing the data center's overall capacity factor. Finally, the research results are used to evaluate the chance of a baseload power demand boom if the IEA high-case scenario unfolds [5].

The remainder of this paper is organized as follows: Section II presents the load profiles of two representative data center types to establish how conventional infrastructure is utilized. Then, the global impact of AI energy use is discussed in terms of existing and predicted AI developments in Section III. Finally, Section IV develops basic economic models of AI energy use to evaluate the economic incentives for either baseload or load-shifting operations before the paper is critically discussed and concluded in Section V.

II. DATA CENTER LOAD CHARACTERIZATION

Fig. 2 depicts a generic power supply sketch for a data center, including both IT-related and cooling-related power demands. However, limited information is available on the power demand profiles of conventional data centers. Fig. 3 and Table 1 describe the mean daily load profiles of two representative types of data centers in the US [12]. Specifically, these include a flat-load data center, which exhibits a high load factor, and a mixed-load data center, which is characterized by load fluctuations that primarily occur during working hours. Although the load characteristics of both types are generally stable and predictable, they exhibit notable seasonal variations. The cooling system has a higher efficiency under colder conditions, implying that the data centers use more power during warmer summer periods. Nevertheless, using adiabatic cooling systems can lead to fewer seasonal variations, even though achieving absolute weather-independence with land-based surface-mounted installations is generally challenging.

The load profiles of flat-load data centers, as shown in Fig. 3-(a), typically have a mean daily load factor[1] of $\geq$ 99 %. Their consistent load is due to auxiliary loads, such as office space consumption, which accounts for a small portion of the total load. As a result, energy consumption showed little correlation with weather or time of day. In contrast, Fig. 3-(b) depicts a representative load profile of a mixed-use data center, where the mean daily capacity factor is $\geq$ 80 %. The peak loads occurring in the middle of the day were similar to those in commercial buildings.

[1] Load factor measures the ratio of the average load to the peak load in a system over a specific period, indicating how evenly the power is used, while the capacity factor assesses the facility's utilization compared to its installed capacity.

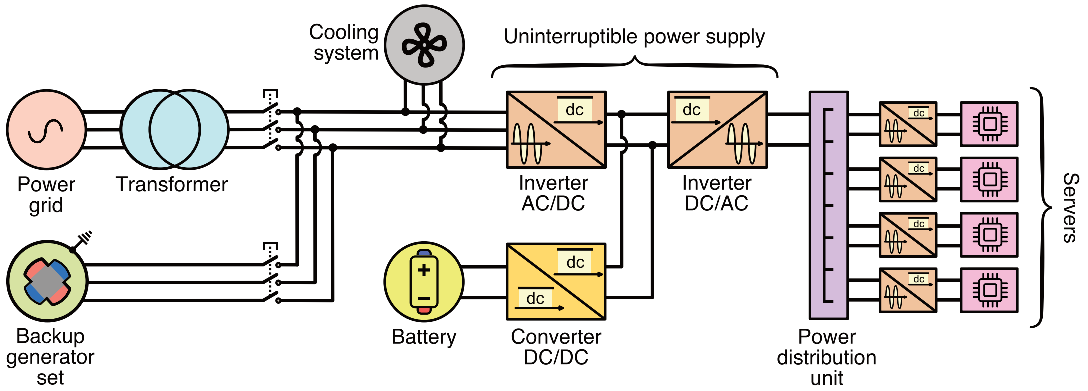

FIGURE 2. Generic power supply schematics for a data center.

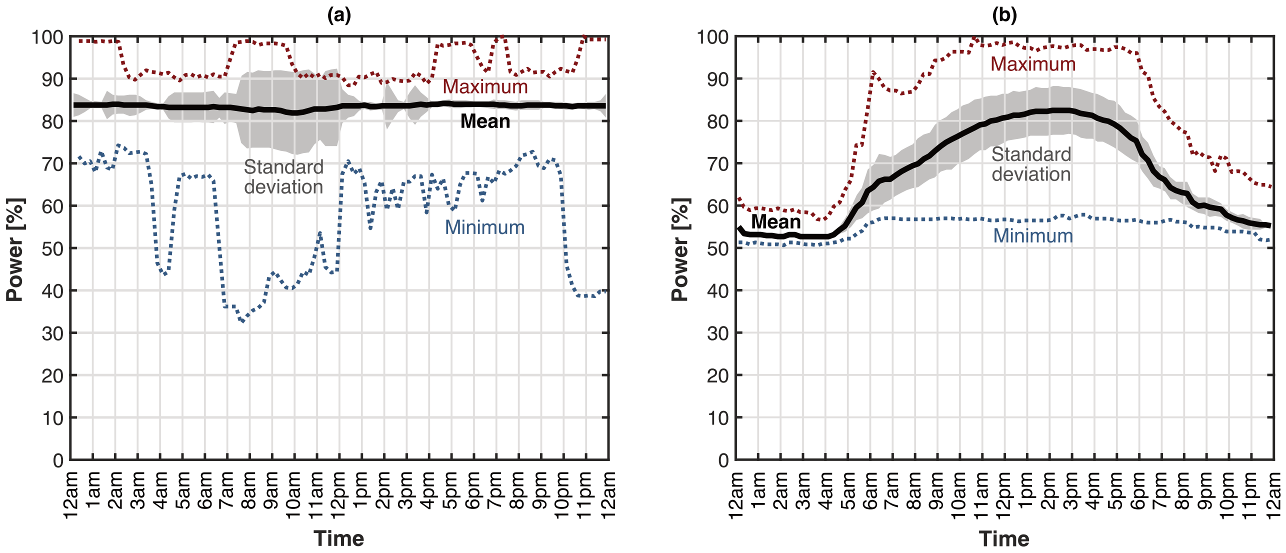

FIGURE 3. Whole-year-averaged load profiles for US data centers based on representative examples [12]. (a) Flat load. (b) Mixed-use load.

In a general sense, Fig. 3 highlights that a significant portion of the data center's power demand is based on a firm, baseload electricity supply around the clock, 24 hours per day, and 365 days per year. However, the mixed-use data center's office-related daily load variations could match well with the variable energy supply of solar power, although this is seasonally dependent. Nonetheless, a data center is a mission-critical, energy-conscious facility with a fundamental requirement of never losing power [19]. Moreover, time is a critical resource during the process of AI model training, and AI data centers are capital-intensive infrastructures that require high utilization to maximize the return on investment. As a result, maximizing their utilization rates is becoming increasingly important for cost-effectiveness. Historically, data centers have been equipped with backup diesel generator sets to secure power reliability and be fully redundant; they also have two independent power grid connections. Table 2 lists the reliability standards of different data center classifications and capital intensities of different reliability levels.

TABLE 1. Characterization of US data center load profiles in Fig. 3 .

	Flat load	Mixed load
Yearly load factor	83.44 %	67.28 %
Mean daily load factor	99.16 %	81.57 %

TABLE 2. Reliability classifications and cost estimations of different data center standards [20], [21].

Tier	Description	Availability	Annual downtime	Overnight cost
I	Basic	99.671 %	28.8 h	$11 500/kW
II	Redundant	99.749 %	22.0 h	$12 500/kW
III	Reliable	99.982 %	1.6 h	$23 000/kW
IV	Fault-tolerant	99.995 %	26.3 min	$25 000/kW

III. IMPACT OF AI USE ON GLOBAL ENERGY USE

The use of AI models is predicted to have a significant impact on global energy consumption in the near future [22]. However, others have argued that some earlier AI claims have been exaggerated [11]. Table 3 shows that the power use per request could increase by as much as 2863 % if the world is transitioning from conventional search services such as Google search to an AI-informed Google search. Considering that 9 billion searches daily would annually create an additional electricity demand of nearly 29.22 TWh or a mean power demand of 3.33 GW, this would be about one-twentieth of the existing data center demand according to IEA reporting. However, another way to implicitly forecast the overall AI power demand is to examine the developments in the AI server market to estimate the power demand required by the future sales and delivery of these infrastructures.

Graphics processing units (GPUs) were originally intended for graphics processing but are now relevant to AI for their ability to parallelize processing tasks and perform rapid tensor operations, which are crucial for AI algorithms. Table 4 highlights the dominance of NVIDIA in the AI server market, where their GPU platforms claim to have the highest computing efficiency. NVIDIA's estimated market share is currently 95 % [5].

The main challenge associated with the theoretical performance metrics of GPU chips, as shown in Table 4, is to bridge them with the real performance of actual AI models. In practice, achieving high GPU utilization requires tasks that are inherently suited for parallel processing, optimized memory management to reduce bottlenecks, and the software using the GPU to be carefully designed to keep all parts of the GPU busy while managing overhead and hardware constraints. Consequently, real-world applications can exhibit GPU utilization as low as 2 % to 10 % [8]. However, certain AI models, such as large language models (LLMs), are well suited for achieving high GPU utilization.

In 2023, NVIDIA is expected to have delivered 100 000 AI server units [6]. Assuming an equal share of NVIDIA DGX$^{\text{TM}}$ A100 and H100 implies a peak power demand from this delivery of 0.84 GW in 2023 alone. Assuming an ideal 100 % load factor and a power usage effectiveness (PUE) factor of 1.12 [23], this implies that 8.2 TWh was added to the new annual consumption as a result of NVIDIAs 2023 global deliveries. By 2027, some have projected that NVIDIA could supply as much as 1 500 000 AI server units [6]. Assuming that it is dominated by the latest product in the DGX series (NVIDIA DGX$^{\text{TM}}$ B200), it would imply as much as 21.45 GW in added peak capacity per year starting from 2027. If fully utilized, it would draw as much as 210.6 TWh of added electricity use per year, assuming a 1.12 PUE. This alone is close to the IEA's high case scenario of 230 TWh in added consumption per year (see Fig. 1), which also covers traditional data centers and cryptocurrencies.

IV. LEVELIZED COST OF COMPUTING

In general, data centers are characterized by a capital-intensive infrastructure. Table 5 lists key economic parameters for estimating the overall cost of computation. The assumed power usage effectiveness (PUE) is low, representing colder regions (e.g., Norway). An economic lifetime of 15 years could be considered optimistic, but the operational cost also includes the replacement of old GPUs. The relatively high discount rate of 10 % can be justified by this short time horizon. The levelized cost of computing (LCOC), inspired by LCOE [44], can be formulated according to eq. (1), which includes capital expenditure (CAPEX), operating expenditure (OPEX), power consumption, and the total amount of computation during the infrastructure lifetime.

TABLE 3. Average power demand per request for different services.

Service	Power use	Increase	Ref.
Conventional Google search	0.30 Wh	–	[24]
ChatGPT	2.89 Wh	+853 %	[22]
BLOOM	3.96 Wh	+1220 %	[25]
AI-powered Google search	8.89 Wh	+2863 %	[22], [26]

$$\text{LCOC} = \frac{c + \sum_{i=0}^{n-1} \frac{\alpha pkt + dc}{(1+r)^i}}{\sum_{i=0}^{n-1} \frac{\gamma kt}{(1+r)^i}} \tag{1}$$

Note that eq. (1) uses the electricity cost from the average electricity price (p) when the data center is in operation, which can also be referred to as the average "capture price" of consumption. Thus, the equation considers the value of demand flexibility for utilizing price variations (hourly and seasonal) by mapping any combination of the average (captured) price and load factor. For example, a combination of a low average price and low load factor reflects a plant that operates only at low prices and is idle for the rest of the year.

A fundamental question is what financial incentives are available for data centers to operate at lower load factors to provide demand response to facilitate renewable production, which is also a general approach to utilize price variations in the electricity market. To answer this question, Fig. 4 highlights the sensitivity of both the load factor and the electricity price to computing costs predicted by eq. (1). This indicates that reduced utilization is a more significant cost driver than the average price of electricity. One critical assumption in the calculation is the computing efficiency, which varies significantly between GPU platforms, as shown in Table 4. These are also theoretical performances if the AI servers are fully utilized in their applications. In Fig. 4, a moderate computing efficiency of 0.1 PFLOPs/kW was assumed. However, this factor is a constant in the LCOC calculation, implying that the relative values of the results remain unchanged, even though it would shift the cost level range of the plots. The costs related to cooling the data center are included in the electricity cost part of eq. (1) through coefficient α, which is the power usage effectiveness (PUE). In this study, a competitive PUE of 1.12 is assumed [23]. However, in some countries with warmer climates, this coefficient will make data centers more sensitive to the cost of energy. Therefore, we added a sensitivity case where the PUE is 1.70, which is representative of a warmer climate. Please

TABLE 4. Performance of different AI data center GPU platforms.

Provider	Platform	Released	Peak power use	Computing performance	Computing efficiency	Ref.
Intel	Data Center GPU Flex 140	2022	0.075 kW	8.000 TFLOPs	0.107 PFLOPs/kW	[27]
Intel	Data Center GPU Flex 170	2022	0.150 kW	16.000 TFLOPs	0.107 PFLOPs/kW	[28]
Intel	Data Center GPU Max 1100	2023	0.300 kW	14.336 TFLOPs	0.048 PFLOPs/kW	[29]
Intel	Data Center GPU Max 1550	2023	0.600 kW	29.491 TFLOPs	0.049 PFLOPs/kW	[30]
AMD	Instinct$^{\text{TM}}$ MI300X	2023	0.750 kW	164.300 TFLOPs	0.209 PFLOPs/kW	[31]
NVIDIA	Tesla T4	2018	0.070 kW	8.100 TFLOPs	0.116 PFLOPs/kW	[32]
NVIDIA	Tesla V100	2017	0.300 kW	15.700 TFLOPs	0.052 PFLOPs/kW	[33]
NVIDIA	DGX$^{\text{TM}}$ A100	2022	6.500 kW	5.000 PFLOPs	0.769 PFLOPs/kW	[34]
NVIDIA	DGX$^{\text{TM}}$ H100	2022	10.200 kW	32.000 PFLOPs	3.137 PFLOPs/kW	[35]
NVIDIA	DGX$^{\text{TM}}$ B200	2024	14.300 kW	72.000 PFLOPs	5.035 PFLOPs/kW	[36]
NVIDIA	GB200 NVL72	2024	120.000 kW	720.000 PFLOPs	6.000 PFLOPs/kW	[37]

TABLE 5. Baseline cost assumptions for data center economic analysis.

Description	Symbol	Value	Ref.
CAPEX	c	\$10 000/kW	[38]–[40]
OPEX	d	10 %	[41]
PUE	α	1.12	[23]
WACC	r	10 %	[42]
Lifetime	n	15	[43]

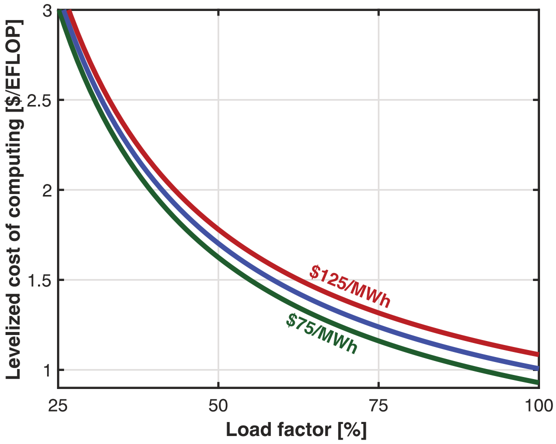

FIGURE 4. Sensitivity to load factor and average price of electricity on the levelized cost of computing (LCOC) using baseline cost assumptions in Table 5 and eq. (1). A 0.1 PFLOPs/kW computing efficiency is assumed, which corresponds with the median value of data center GPU platforms in Table 4. The load factor is varied between 25 and 100 % and the investigated electricity prices are \$75/MWh, \$100/MWh, and \$125/MWh, where the latter is close to the 2023 wholesale electricity price of \$127.2/MWh in the US [45].

note that if fossil fuels constitute a significant portion of the data center's power supply, carbon pricing must be added to eq. (1).

To further explore cost sensitivities, Fig. 5 provides a two-dimensional (2-D) contour plot of the LCOC with respect to both the load factor and the electricity price. The green contour line has an LCOC of \$1.01/EFLOP and a baseload electricity price of \$100/MWh at 100 % load factor. The same LCOC is achieved at a 69.10 % load factor with a \$0/MWh electricity price. This means that electricity would have to be free if the data center were to reduce its load factor to 69.10 % while maintaining the same LCOC. This is because the capital expenditures and operational costs in the data center sector are much higher than the costs of electricity, which are different from other power-intensive industries that are very sensitive to the average price of electricity.

Table 6 summarizes all the zero electricity price load factors that match the contour plots in Fig. 5. High-tier data centers with even higher CAPEX values (see Table 2) would shift the contours in Fig. 5 further to the right. Achieving a high-reliability classification primarily depends on the reliability of the main power source. Therefore, a higher average price of electricity may be justified for a more reliable power supply. Hence, a separate investment in local energy solutions with sufficient reliability next to data centers could transform a low-CAPEX tier-I data center with performance comparable to a high-CAPEX tier-IV data center without the need to acquire backup power from in-house diesel generator sets or other reliable solutions inside the data center (see Fig. 2).

The contour plots in Fig. 5 highlight that a lower baseload electricity price provides less room to reduce the load factor and still be economically competitive with baseload operation. In terms of cost, a lower load factor should be justified by capturing a lower average price of electricity at the expense of lower utilization. This economic window of opportunity shrinks as the baseload electricity prices become more competitive. Even though there is an economic opportunity available, the potential load factors are still high, implying that data centers in the future will very likely be consumers dominated by baseload operation at high load factors. However, data centers can sell system-bearing services to grid operators to economically justify lower load factors. In addition, other remuneration mechanisms for providing demand response could offer further financial incentives, such as workload scheduling. Nonetheless, the cost-effectiveness of such services might be higher when sourced from entities other than capital-intensive, high-operating-cost facilities such as data centers.

When older GPUs are replaced with newer models with higher computational efficiency, the LCOC improves

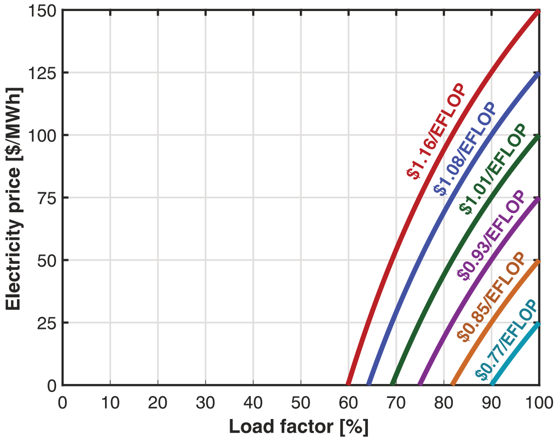

FIGURE 5. 2-D contour plots of the levelized cost of computation (LCOC) with respect to both load factor and average price of electricity using baseline cost assumptions in Table 5 and eq. (1). A 0.1 PFLOPs/kW computing efficiency is assumed, which is the median value of data center GPU platforms in Table 4. The zero-crossings are specified in Table 6.

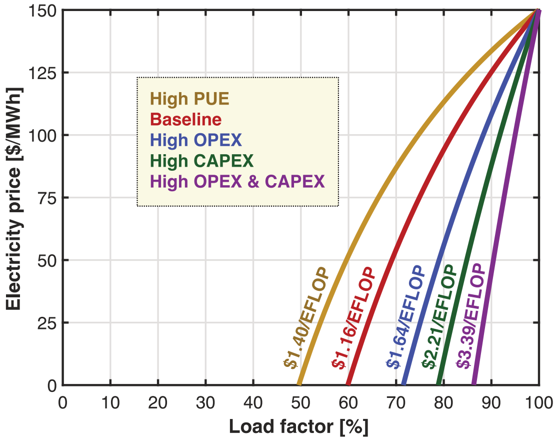

FIGURE 6. Sensitivity to different input parameters for the 2-D contour plots of the levelized cost of computation (LCOC) with a baseload electricity price of $150/MWh and the same baseline parameters as in Fig. 5. Sensitivity cases include high PUE (1.70) relevant to warmer climates, high OPEX (25 % of CAPEX annually), high CAPEX ($25 000/kW), and both high OPEX and high CAPEX.

TABLE 6. Zero electricity price crossings in the contour plots of Fig. 5 using the baseline cost assumptions in Table 5.

Price level	Baseload electricity price	Levelized cost of computation	Zero electricity price load factor
#1	$150/MWh	$1.16/EFLOP	59.85 %
#2	$125/MWh	$1.08/EFLOP	64.14 %
#3	$100/MWh	$1.01/EFLOP	69.10 %
#4	$75/MWh	$0.93/EFLOP	74.88 %
#5	$50/MWh	$0.85/EFLOP	81.72 %
#6	$25/MWh	$0.77/EFLOP	89.94 %
#7	$0/MWh	$0.70/EFLOP	100.00 %

TABLE 7. Zero electricity price crossings in the contour plots of Fig. 6, assuming a baseload electricity price of $150/MWh.

	Levelized cost of computation	Zero electricity price load factor
High PUE	$1.40/EFLOP	49.55 %
Baseline	$1.16/EFLOP	59.85 %
High OPEX	$1.64/EFLOP	71.50 %
High CAPEX	$2.21/EFLOP	78.84 %
High OPEX & CAPEX	$3.39/EFLOP	86.26 %

accordingly. Consequently, profits can be higher than expected when the first investment decision is made. In this way, data centers may justify a higher electricity cost when entering into a long-term power purchase agreement (PPA) as long as power reliability can be assured.

To highlight PUE, CAPEX, and OPEX sensitivities, Fig. 6 shows the impact of higher PUE, OPEX, and CAPEX. High CAPEX and OPEX levels were set to typical values for tier-IV data centers [20], [21]. These results reinforce the findings in Fig. 5, as the contour curves are shifted further to the right and limit incentives for flexible operation. Nevertheless, a higher PUE level shifts the contour to the left, increasing the incentives for flexibility, albeit at a 21 % higher LCOC. This highlights the value of locating data centers in colder climates to enable more efficient cooling and, thereby, a lower PUE. The zero-crossings of Fig. 6 are given in Table 7.

V. DISCUSSION AND CONCLUSION

This paper reviews recent expected developments in AI energy use and presents an economic model to evaluate the load profile of AI data centers. In general, firm power availability is found to have high economic incentives, and electricity price levels and price variations have a lower role in ensuring the competitiveness of the data center infrastructure.

As mentioned in the literature review of this paper, there have been some investigations and proposals to load-shift data centers to make them easier to directly integrate with variable renewable energy. The data centers could be running computations during the day when the sun is shining or during multiple days of high wind power output. However, the cost of AI training is particularly sensitive to the load factor, which limits the incentives for load shifting. There might also be issues due to bottlenecks in the power grid transmission capacity, which makes it challenging to deploy data centers far away from the best renewable resources.

In the short term, the expansion of AI infrastructure could boost the deployment of classical baseload generation facilities such as coal and combined-cycle natural gas. However, gas power plants are limited to regions with competitive gas prices. In the long term, carbon prices and the depletion of fossil fuel resources can limit their competitiveness with other technologies. In this case, clean baseload alternatives will take over. They can be provided by nuclear energy (e.g., centralized large reactors or pools of local small modular reactors), variable renewables with long-term energy storage, and natural gas and coal equipped with carbon capture and storage (CCS) technology.

In future research, the composition of AI data centers could be studied in more detail, separating out the cost of GPUs and CPUs and their depreciation to better understand how AI data center composition and their use cases impact the CAPEX and OPEX terms differently from the tier level classifications considered in this study.

REFERENCES

[1] R. Verdecchia, J. Sallou, and L. Cruz, ''A systematic review of green AI,'' *WIREs Data Mining Knowl. Discovery*, vol. 13, no. 4, p. e1507, Jul. 2023.

[2] IEA. (2023). *Data Centres and Data Transmission Networks*. [Online]. Available: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

[3] N. S. Malik. (2024). *AI Is Exploding Data Center Energy Use. A Google-Created Technique May Help*. Bloomberg. [Online]. Available: https://www.bloomberg.com/news/articles/2024-02-25/ai-increases-data-center-energy-use-google-pioneered-technique-could-help

[4] Economist. (2024). *Big Tech's Great AI Power Grab*. [Online]. Available: https://www.economist.com/business/2024/05/05/big-techs-great-ai-power-grab

[5] IEA. (2024). *Electricity 2024—Analysis and Forecast to 2026*. [Online]. Available: https://www.iea.org/reports/electricity-2024

[6] E. Bary. (2023). *Nvidia is 'Dominating' and Could Unlock $300 Billion in AI Revenue by 2027, Analyst Says*. MarketWatch. [Online]. Available: https://www.marketwatch.com/story/nvidia-is-dominating-and-could-unlock-300-billion-in-ai-revenue-by-2027-analyst-says-915935c0

[7] ScienceDaily. (2023). *Powering AI Could Use as Much Electricity as a Small Country*. [Online]. Available: https://www.sciencedaily.com/releases/2023/10/231010133607.htm

[8] R. West. (2024). *Energy and AI: The Power and the Glory?* Thunder Said Energy. [Online]. Available: https://thundersaidenergy.com/2024/04/04/energy-and-ai-the-power-and-the-glory/

[9] 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.

[10] M. Koot and F. Wijnhoven, ''Usage impact on data center electricity needs: A system dynamic forecasting model,'' *Appl. Energy*, vol. 291, Jun. 2021, Art. no. 116798.

[11] D. Castro. (2024). *Rethinking Concerns About AI's Energy Use*. Centre for Data Innov. [Online]. Available: https://www2.datainnovation.org/2024-ai-energy-use.pdf

[12] G. Ghatikar, M. A. Piette, S. Fujita, A. McKane, J. H. Dudley, A. Radspieler, K. C. Mares, and D. Shroyer, ''Demand response and open automated demand response opportunities for data centers,'' Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep. LBNL-3047E, 2010. [Online]. Available: https://eta-publications.lbl.gov/sites/default/files/demand_response_and_open_automated_demand_response_opportunities_for_data_centers_lbnl-3047e.pdf

[13] A. Krioukov, C. Goebel, S. Alspaugh, Y. Chen, D. E. Culler, and R. H. Katz, ''Integrating renewable energy using data analytics systems: Challenges and opportunities,'' *IEEE Data Eng. Bull.*, vol. 34, no. 1, pp. 3–11, Mar. 2011.

[14] Ì. Goiri, M. E. Haque, K. Le, R. Beauchea, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, ''Matching renewable energy supply and demand in green datacenters,'' *Ad Hoc Netw.*, vol. 25, pp. 520–534, Feb. 2015.

[15] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, ''Renewable and cooling aware workload management for sustainable data centers,'' in *Proc. 12th ACM SIGMETRICS/PERFORMANCE Joint Int. Conf. Meas. Model. Comput. Syst.*, Jun. 2012, pp. 175–186.

[16] C. Chen, B. He, and X. Tang, ''Green-aware workload scheduling in geographically distributed data centers,'' in *4th IEEE Int. Conf. Cloud Comput. Technol. Sci. Proc.*, Dec. 2012, pp. 82–89.

[17] Z. Liu, M. Lin, A. Wierman, S. Low, and L. L. H. Andrew, ''Greening geographical load balancing,'' *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 657–671, Apr. 2015.

[18] A. Nadjaran Toosi, C. Qu, M. D. de Assunção, and R. Buyya, ''Renewable-aware geographical load balancing of web applications for sustainable data centers,'' *J. Netw. Comput. Appl.*, vol. 83, pp. 155–168, Apr. 2017.

[19] B. Tita. (2024). *For AI, a Few Seconds of Power Becomes a Booming Business*. Wall Street J. [Online]. Available: https://www.wsj.com/business/energy-oil/for-ai-a-few-seconds-of-power-becomes-a-booming-business-c16cb626

[20] M. Stansberry. (2021). *Explaining the Uptime Institute's Tier Classification System*. Uptime Inst. [Online]. Available: https://journal.uptimeinstitute.com/explaining-uptime-institutes-tier-classification-system/

[21] KIO. (2019). *Costs of a Data Center*. [Online]. Available: https://www.kio.tech/en-us/blog/data-center/costs-of-a-data-center

[22] A. de Vries, ''The growing energy footprint of artificial intelligence,'' *Joule*, vol. 7, no. 10, pp. 2191–2194, Oct. 2023.

[23] S. Saha, J. Sarkar, A. Dwivedi, N. Dwivedi, A. M. Narasimhamurthy, and R. Roy, ''A novel revenue optimization model to address the operation and maintenance cost of a data center,'' *J. Cloud Comput.*, vol. 5, no. 1, pp. 1–23, Dec. 2016.

[24] U. Hölzle. (2009). *Powering a Google Search*. Google. [Online]. Available: https://googleblog.blogspot.com/2009/01/powering-google-search.html

[25] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, ''Estimating the carbon footprint of bloom, a 176 b parameter language model,'' *J. Mach. Learn. Res.*, vol. 24, no. 253, pp. 1–15, 2023.

[26] D. Patel and A. Ahmad. (2023). *The Inference Cost of Search Disruption—Large Language Model Cost Analysis*. SemiAnalysis. [Online]. Available: https://www.semianalysis.com/p/the-inference-cost-of-search-disruption

[27] Intel. (2022). *Intel Data Center GPU Flex 140*. [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/230020/intel-data-center-gpu-flex-140/specifications.html

[28] Intel. (2022). *Intel Data Center GPU Flex 170*. [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/230019/intel-data-center-gpu-flex-170/specifications.html

[29] Intel. (2023). *Intel Data Center GPU Max 1100*. [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/232876/intel-data-center-gpu-max-1100/specifications.html

[30] Intel. (2023). *Intel Data Center GPU Max 1550*. [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/232873/intel-data-center-gpu-max-1550/specifications.html

[31] AMD. (2023). *AMD Instinct MI300X Accelerators*. [Online]. Available: https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html

[32] NVIDIA. (2018). *NVIDIA T4—Flexible Design, Extraordinary Performance*. [Online]. Available: https://www.nvidia.com/en-us/data-center/tesla-t4/

[33] NVIDIA. (2017). *NVIDIA Tesla V100—The First Tensor Core GPU*. [Online]. Available: https://www.nvidia.com/en-gb/data-center/tesla-v100/

[34] NVIDIA. (2022). *NVIDIA DGX A100—The Universal System for AI Infrastructure*. [Online]. Available: https://images.nvidia.com/aem-dam/Solutions/Data-Center/nvidia-dgx-a100-datasheet.pdf

[35] NVIDIA. (2022). *NVIDIA DGX H100—The Gold Standard for AI Infrastructure*. [Online]. Available: https://resources.nvidia.com/en-us-dgx-systems/ai-enterprise-dgx

[36] NVIDIA. (2024). *NVIDIA DGX B200—A Unified AI Platform for Training, Fine-Tuning, and Inference*. [Online]. Available: https://resources.nvidia.com/en-us-dgx-systems/dgx-b200-datasheet

[37] NVIDIA. (2024). *NVIDIA GB200 NVL72—Powering the New Era of Computing*. [Online]. Available: https://www.nvidia.com/en-us/data-center/gb200-nvl72/

[38] K. Pilz and L. Heim, ''Compute at scale: A broad investigation into the data center industry,'' 2023, *arXiv:2311.02651*.

[39] L. Repta. (2020). *Itasca Country Club Data Center—Fiscal Impact Analysis*. Itasca. [Online]. Available: https://www.itasca.com/DocumentCenter/View/9389/16—Economic-Impact-Memo

[40] R. K. Hill. (2020). *Project Oasis—Market Analysis for Data Center Investment in Southwest Virginia*. OnPoint Develop. Strategies. [Online]. Available: https://static1.squarespace.com/static/5f0a2bd5c3354d1c75ad855e/t/5f78 a1c787afb65ed3150ff1/1601741260417/Project+Oasis+Final+Report

[41] T. Day and N. D. Pham. (2017). *Data Centers: Jobs & Opportunities in Communities Nationwide*. U.S. Chamber Commerce. [Online]. Available: https://www.uschamber.com/technology/data-centers-jobs-opportunities-communities-nationwide

[42] M. Norris. (2021). *Data Centres: The Concrete Behind 'Cloud Computing'*. FT Adviser. [Online]. Available: https://www.ftadviser.com/Partner-Contents87/2021/05/27/Gravis-Data-Centres-The-Concrete-Behind-Cloud-Computing

[43] J. Koomey, K. Brill, P. Turner, J. Stanley, and B. Taylor, "A simple model for determining true total cost of ownership for data centers," Uptime Inst., New York, NY, USA, White Paper TUI3011B, 2007. [Online]. Available: https://www.missioncriticalmagazine.com/ext/resources/MC/Home/Files/PDFs/(TUI3011B)SimpleModelDetermingTrueTCO.pdf

[44] W. Shen, X. Chen, J. Qiu, J. A. Hayward, S. Sayeef, P. Osman, K. Meng, and Z. Y. Dong, "A comprehensive review of variable renewable energy levelized cost of electricity," *Renew. Sustain. Energy Rev.*, vol. 133, Nov. 2020, Art. no. 110301.

[45] B. Alves. (2024). *Average Retail Electricity Prices in the United States From 1990 to 2023*. Statista. [Online]. Available: https://www.statista.com/statistics/183700/us-average-retail-electricity-price-since-1990/

JONAS KRISTIANSEN NØLAND (Senior Member, IEEE) was born in Drammen, Norway, in 1988. He received the M.Sc. degree in electric power engineering from the Chalmers University of Technology, Gothenburg, Sweden, in 2013, and the Ph.D. degree in engineering physics from Uppsala University, Uppsala, Sweden, in 2017. Since 2018, he has been an Associate Professor with the Department of Electric Energy, Norwegian University of Science and Technology (NTNU). He is currently involved in the NERES Project, exploring nuclear energy's role in a renewable energy system. He has been serving as the Chair for the IEEE Power and Energy Society (PES) Norwegian Chapter, since 2022. He serves as an Associate Editor for IEEE Transactions on Energy Conversion and IEEE Transactions on Industrial Electronics.

MARTIN HJELMELAND (Member, IEEE) received the M.Sc. and Ph.D. degrees in electric power engineering from Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2015 and 2019, respectively. Earlier in his career, he focused on medium-term hydropower scheduling in multiple markets. He is currently a Postdoctoral Researcher with the Department of Electric Energy (IEL), NTNU, where he is a member with the "Electricity Markets and Energy Systems Planning" (EMESP) Research Group. He is also involved in the strategic research project "Nuclear Energy's role in a Renewable Energy System (NERES)." His research interests include the integration of low-carbon energy sources in the energy systems, with a particular focus on the role of nuclear energy.

MAGNUS KORPÅS (Member, IEEE) received the M.S. degree in theoretical physics and the Ph.D. degree in electric power engineering on the topic of optimizing the use of energy storage for distributed wind energy in the power market from Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 1998 and 2004, respectively.

From 2004 to 2006, he was a Postdoctoral Fellow with NTNU. He held several research positions (a Scientist, the Manager, and the Director) with SINTEF Energy, Trondheim, from 2006 to 2014, where he was the Leader and an active participant in several large energy research projects at national and European levels. In 2014, he became a Full Professor with the Department of Electric Power Engineering, NTNU, where he established the Electricity Markets and Energy System Planning Research Group. He was a Visiting Researcher with the MIT Laboratory for Information and Decision Systems (LIDS), from 2018 to 2019. He is currently the Leader of the Scientific Committee, the work package on flexible resources in the power system with the Centre for Intelligent Electricity Distribution (CINELDI), and the work package markets in the IEA Wind Task 25 (design and operation of energy systems with large amounts of variable renewable generation).