Data Center Developments for Flexible Generation Dispatch, Advanced Infrastructure, and Ultra-Fast Digital Twins

Grant M. Fischer, Rosemary E. Alden, Donovin D. Lewis, Aron Patrick[1], and Dan M. Ionel

SPARK Laboratory, Stanley and Karen Pigman College of Engineering, University of Kentucky, Lexington, KY, USA
[1]PPL Corporation, Allentown, PA
grant.fischer@uky.edu, rosemary.alden@uky.edu, donovin.lewis@uky.edu, aronpatrick@gmail.com, dan.ionel@ieee.org

***Abstract*—The rapid advancement and widespread integration of artificial intelligence (AI) is driving demand for unprecedented deployment of power-intensive computational infrastructure, including multi-megawatt data centers with the potential for facilities with gigawatt-scale capacity in the near future. In this paper, load growth projections for the US are reviewed, and an example energy dispatch solution considering a mixed energy portfolio with flexible, renewable, distributed, and load-based generation is employed. The brief technology review included in the paper covers aspects of electric power, cooling, and computational infrastructures. The concept of a data center digital twin for transient load, grid interaction, and hybrid energy dispatch studies is described and proposed for implementation using a power hardware-in-the-loop test bench. For the transients associated with the typical data center loads, a generative adversarial neural (GAN) model is proposed and employed to synthetically produce GPU load profiles at the rack level, which are aggregated to emulate large-scale data center behavior.**

***Index Terms*—artificial intelligence (AI), transient load, generative adversarial neural network (GAN), power hardware-in-the-loop (PHIL)**

I. Introduction

The growth of internet traffic has historically driven the expansion of data center computational infrastructure; however, from 2010 to 2020, the energy needed to supply these servers remained relatively stable due to improvements in cooling efficiency and consolidation through cloud computing [1]. New demand for internet and computational services, especially artificial intelligence (AI) applications, is accelerating the construction of new data centers at a rate that exceeds the pace of efficiency improvements, resulting in substantial growth in overall electrical load.

Demand for AI computing capacity is expected to require more than a trillion dollars in investments in data center infrastructure by 2030, driven especially by large hyperscale tech companies and competition between governments around the world [2, 3]. This new investment, primarily concentrated in the United States, is cause for concern for the electric power grid amid the ongoing pressure from other industries, including efforts for electrification and industrial onshoring in the United States [4]. Load forecasting is essential for maintaining grid reliability and resilience, and a deeper understanding of data center load growth is required to reduce uncertainty in future grid planning.

The design of data center architectures that incorporate efficient cooling systems and renewable generation is important for supporting energy-efficient operation and effective integration of renewable resources. Furthermore, AI workloads in data centers can induce large-scale power transients, resulting in megawatt-level fluctuations in demand [5]. Accurate modeling of these effects is needed for developing data center architectures with reliable power infrastructures capable of accommodating such dynamic behaviors.

Synthetic generation of training load power offers a path to estimating the range and volatility of AI training at scale. Data-driven methods such as generative AI have been limited with few public datasets with sufficient power profiles. Recent efforts to standardize evaluation of large language learning models (LLM) training energy efficiency across systems and power levels from over 60 systems have provided a new resource, MLPerf Power [6]. Within this set, there are over 1,400 power vs. time LLM training curves that are employed in this first of a kind data-driven effort to synthetically generate aggregated data center power swings at high resolution.

This paper reviews US data center load growth projections and includes an example load-based generation dispatch for incorporating diverse, flexible, and renewable generation resources. A review of data center architecture, including cooling, computational, and power support infrastructures, provides a summary of the systems needed for facility operation and for developing a data center digital twin. A generative adversarial neural (GAN) model is employed for generating synthetic AI workload transient power profiles. Synthetic and real data are aggregated in maximum overlap and random overlap scenarios showing large transients that may occur with unmanaged LLM training workloads.

II. Load Growth Projections and Generation Dispatch

Data centers are projected to contribute to significant electrical load growth as a result of the widespread adoption of artificial intelligence. The IEA estimates that data centers accounted for 415TWh of the world's electricity use in 2024,

Table I: US Data Center Load Growth Forecasts.

Organization	Forecasted Timeline	% Demand Growth/Year	% US Load Initial	% US Load End
EPRI (High) [1]	2023–2030	10%	4.0%	6.8%
EPRI (Higher) [1]	2023–2030	15%	4.0%	9.1%
LBNL (Low) [7]	2023–2028	13%	4.4%	6.7%
LBNL (High) [7]	2023–2028	23%	4.4%	12.0%
IEA [8]	2024–2030	22%[1]	4.0%[2]	—
Boston Consulting Group [3, 9]	2022–2030	16%[3]	2.5%	7.5%

[1] Calculated growth per year from 130% projected cumulative growth over the 2024–2030 forecasted time frame.
[2] The 4% initial US load is the minimum IEA calculation.
[3] Boston Consulting Group includes a minimum year-over-year growth of 12%.

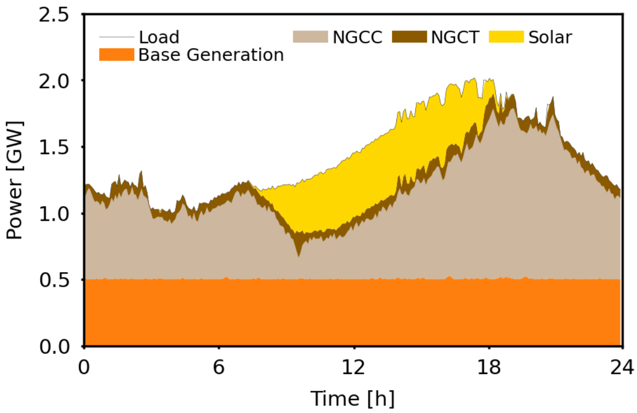


Fig. 1: Generation dispatch for example day including solar, base generation, and natural gas generation resources. Base generation provides relatively continuous power with natural gas combined cycle and combustion turbine generation ramping to meet demand and flex with solar generation.

and that that the US, which makes up 45% of global demand, will account for half of the growth contributing to the projected 945TWh of demand by 2030 [8]. This forecast suggests data center load growth as one of the fastest growing electrical loads in the world, and uncertainty remains around how much of the planned capacity increase will be realized.

Demand growth projections for data centers in the US are summarized in Table I, including annual demand growth rates and the share of total national load by the end of each forecasting period. The projected annual growth rates range from 10% to 23%, highlighting both the high uncertainty of these projections and the rapid increase in expected data center capacity. In all cases, data centers are projected to account for an increasing share of US electricity demand, with most cases predicting more than twice the total demand by 2030. Considering the uncertainty surrounding AI expansion and its supporting infrastructure, some projections include low growth cases, such as EPRI's low scenario for 3.7% annual increase and a 4.6% share of total U.S. load by 2030 [1].

Due to the growth in large-scale data centers and cloud computing, new data center installation can be highly localized and dramatically increase the power demand needed in a short period of time. Loudoun County in northern Virgina, for example, has an operating data center power demand of 5.3GW with a further 6.3GW planned for future development [10]. From 2018 to 2022, the square footage devoted to data centers in Loudoun doubled and is expected to continue to increase [11]. The county is working to install onsite solar and wind, reduce dependence on diesel backup generators, and increase the use of off-site renewables for new data centers.

In a 2024 survey including 22 electric utilities from across the US and 3 international utilities, 23% of responding utilities reported cumulative data center requests greater than or equal to 100% of their peak load, 48% have received connection request for data centers with capacity of at least 1GW, and 26% of utilities have experienced ramp rate issues due to already connected data centers [12]. Increased generation and dispatch control may be needed to support new data center installation, and renewable energy generation may help support growing demand and enhance grid stability through effective energy management and spatially informed integration of distributed renewable resources [13–15].

Example results from a case study employing load-based generation dispatch, described in Lewis et al., is depicted in Fig. 1 with five minute load and solar data in a mixed energy portfolio optimizing flexible thermal generation output [16, 17]. Base generation resources can include relatively slow ramping power plants such as nuclear or coal. Solar and base generation resources supply their available output, while flexible control of natural gas units is applied to meet the remaining demand. This algorithm demonstrates coordinated control of distributed and renewable energy resources, integrating flexible and non-flexible generation to balance load throughout the day.

Future studies can review the potential for data centers to be more than just an independent electrical load for the power grid. The unique qualities of data center operation could allow for spatial-temporal load shifting to run more power intensive operations in periods of high renewable production for flexible collaboration with grid operators [18].

III. Infrastructure and Digital Twin

To improve efficiency and computational capacity, data center infrastructure technologies are trending toward supporting high power density, large-scale data centers. Improvements to support these infrastructures could include advanced cooling methods, continuous and redundant power support, and investment in power intensive parallel processing AI accelerators,

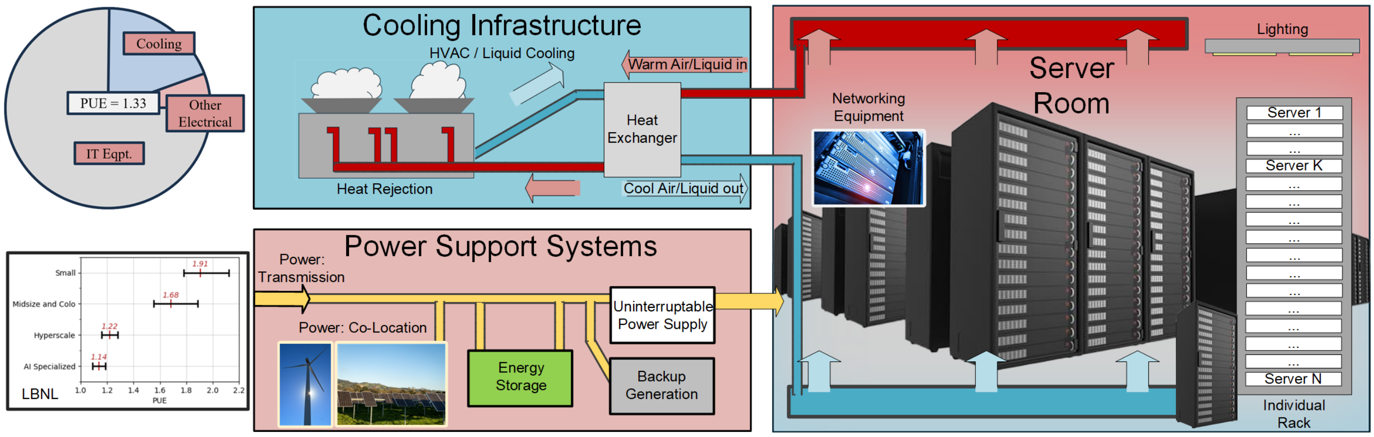


Fig. 2: Data center architecture overview depicting the primary systems devoted to supporting data center operation, and an example PUE chart for infrastructure percent power utilization. Example PUE ranges of typical AI specialized hyperscale, colocation, and small data centers are included to show that hyperscale and AI specialized data centers are far more efficient than the average traditional data center.

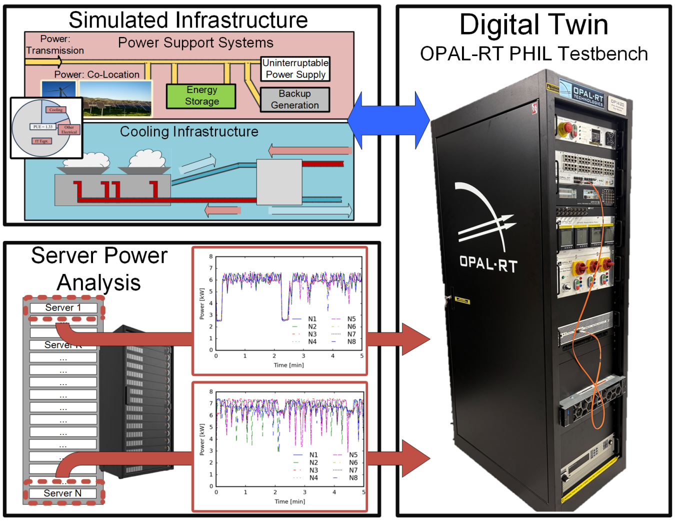


Fig. 3: Proposed simulated infrastructure and power analysis for digital twin utilizing PHIL test bench. Server power analysis includes example power draw profiles of eight computing nodes (N1 - N8) during training of Llama 270B LoRA and GPT-3 models on NVIDIA H100-SXM GPUs.

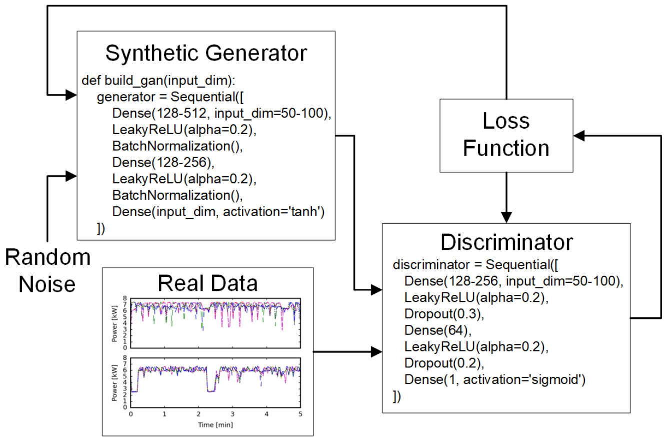
such as GPUs. Cloud computing has enabled a transition from small server facilities to large colocation centers, where computing resources can be rented or purchased, and to hyperscale data centers operated by major tech companies, where these architectural improvements can be implemented on a large-scale for improving overall efficiency.

The power usage effectiveness (PUE) of example data centers is presented as a part of the architecture overview in Fig. 2 with a comparison between advanced data center designs, like hyperscale centers, and small-scale data centers, which are relatively inefficient when comparing PUE [7]. The PUE is calculated by dividing the total power supplied to a data center by the power that is directly utilized for information technology (IT) equipment. Therefore, data center PUE increases when a higher proportion of the total power is utilized by IT equipment. Other electrical systems typically make up the smallest percentage of power use, including those that support lighting, networking, and security.

An example of the typical power and cooling infrastructure needed for data center operation is also shown in Fig. 2. Cooling infrastructure accounts for the second-largest share of power demand on average in data centers with a PUE less than two and can include air conditioners, chillers, economizers, and dry coolers depending on the type of data center [7]. Advancements in cooling system technology have greatly improved energy efficiency over time, and the adoption of direct liquid cooling for IT equipment is an emerging solution that could continue efficiency improvement. The example power delivery infrastructure is designed to ensure continuous operation, with uninterruptible power supplies sustaining short outages and energy storage or backup generators maintaining supply during extended outages or grid disturbances.

Transient and highly variable GPU power profiles during AI training workloads over a five minute interval are shown in Fig. 3 with power measured measured in kW [19]. Significant power fluctuations occur over just a few seconds, including rapid jumps of load with up to a 16% ripple in power draw within these public profiles. Analysis of the transient behavior in data center loads, especially when at the MW/GW scale, is critical for building the grid to support data centers as they continue to make up a growing percentage of the electric power demand.

One solution for this second level analysis is to develop an ultra-fast digital twin utilizing state-of-the-art power hardware-in-the-loop (PHIL) equipment. An ultra-fast digital twin utilizing PHIL could analyze real-time power profiles of GPUs,



```
def build_gan(input_dim):
    generator = Sequential([
        Dense(128-512, input_dim=50-100),
        LeakyReLU(alpha=0.2),
        BatchNormalization(),
        Dense(128-256),
        LeakyReLU(alpha=0.2),
        BatchNormalization(),
        Dense(input_dim, activation='tanh')
    ])
```

```
discriminator = Sequential([
    Dense(128-256, input_dim=50-100),
    LeakyReLU(alpha=0.2),
    Dropout(0.3),
    Dense(64),
    LeakyReLU(alpha=0.2),
    Dropout(0.2),
    Dense(1, activation='sigmoid')
])
```

Fig. 4: A visualization of GAN structure. The generator is tuned through competition with the discriminator to create accurate synthetic data.

Table II: Training performance across different signal lengths and generation settings with accuracy on synthetic signals.

Len. [s]	Quantity	Generated	Epochs	Disc. Accuracy
60	52	50	33	0.66
1701	98	50	36	0.44
158	458	50	34	0.44
		100	49	0.57
58	304	50	27	0.62
		100	48	0.46
59	202	50	20	0.58
		100	36	0.61

simulate data center infrastructure and grid interaction, and apply simulated hybrid energy dispatch to support variable data center loads [20, 21].

IV. Electric Power Estimations for Data Centers through Generative AI

The data for the proposed PHIL testbeds for rack level power transients analysis may be obtained with the data-driven GAN models described in this section. These models may support a first evaluation of scalable synthetic load volatility and of smoothing effects at the aggregate level. A general visualization of GAN structure is included in Fig. 4.

Data availability is a challenge across all deep learning research and especially so with the rapid growth of data centers. Recent efforts by companies and organizations including, Google and Microsoft, have provided new gains for public data with a set of LLM training curves from the following models: DLRM-DCNv2, GPT-3, Llama 2 70B LoRA, ResNet, SSD, 3D U-Net, and BERT [6]. The company SMC included over 1,400 rack level AC power signals of different lengths from 60s to 28mins at second resolution.

The field of synthetic data generation for electric power loads has been established for residential, industrial, and agricultural power profiles with acceptable accuracy scores for vanilla GAN and best scores for Conv1D-WGAN-GP variants [22]. Based on these precedents, vanilla GAN models have been trained on the AC power signals (Table. II). Two cases in-which 50 and 100 new signals were produced in which synthetic signals passed as real, fooling the discriminator between 44-66% of the time across lengths. During training instability occurred intermittently leading to difficulty with consistent discriminator (disc.) accuracy.

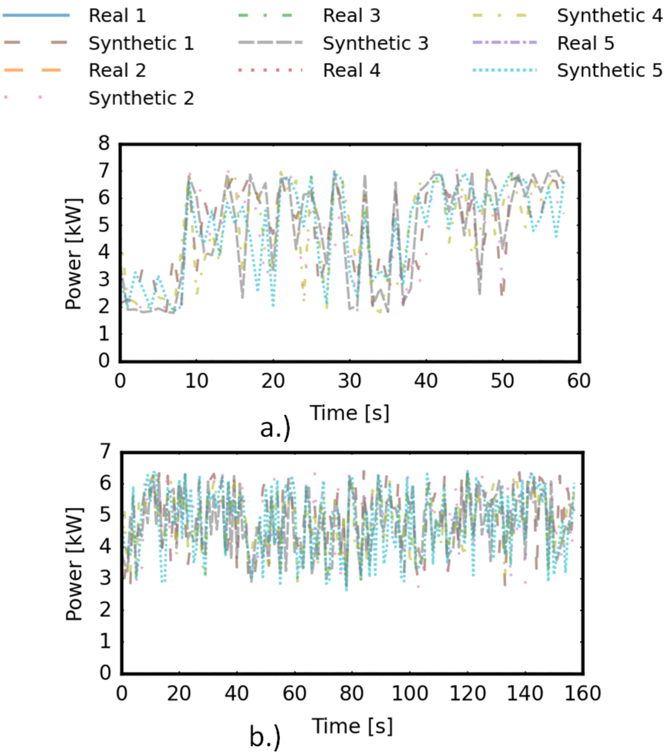


Fig. 5: Example real and GAN generated synthetic data center AC rack power for signals of length 58 and 160s based on the MLPerf online repository.

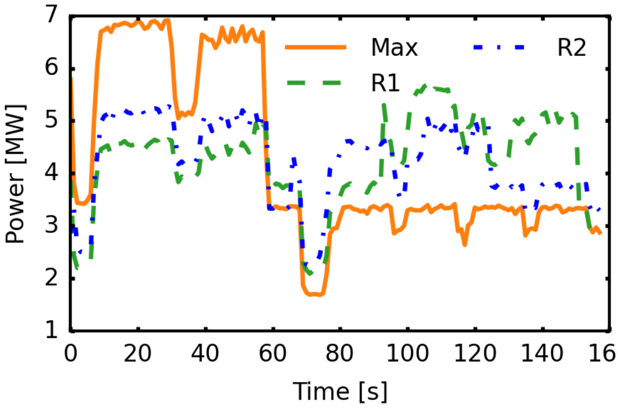


Fig. 6: Aggregate data center load profiles employing the real and synthetic data for training LLMs. A maximum overlap and two random overlap cases show the wide shifts in load that may occur if large training jobs end at the same time.

Algorithm Aggregate load generation of data centers during training from real and synthetic profiles.

```
Require: Set of signals $S = \{s_0, s_1, \ldots, s_N\}$, padding mode $\mathsf{mode} \in \{\mathit{max}, \mathit{random}\}$
Ensure: Aggregated signal $A$
$L_{\text{max}} \leftarrow \max_i \text{length}(s_i)$
Initialize padded signal set $\tilde{S} \leftarrow \emptyset$
for each signal $s_i \in S$ do
    $L_i \leftarrow \text{length}(s_i)$
    $\text{pad}_i \leftarrow L_{\text{max}} - L_i$
    if $\text{pad}_i > 0$ then
        if $\mathsf{mode} = \mathit{max}$ then
            $\tilde{s}_i \leftarrow \text{Pad}(s_i, \text{pad}_i)$ with zeros at the end
        else if $\mathsf{mode} = \mathit{random}$ then
            Sample $k_i \sim \mathcal{U}(0, \text{pad}_i)$
            Pre-pad $s_i$ with $k_i$ zeros
            Post-pad $s_i$ with $\text{pad}_i - k_i$ zeros
            $\tilde{s}_i \leftarrow$ padded signal
        end if
    else
        $\tilde{s}_i \leftarrow s_i$
    end if
    Append $\tilde{s}_i$ to $\tilde{S}$
end for
$A = \sum_{i=0}^{N} \tilde{s}_i \,\Big|\, \tilde{s}_i \in \tilde{S}$
return $A$
```

There are known reasons for this instability in vanilla GAN models, and further inclusions of improved constructions, such as with the Wasserstein GAN with gradient penalty, may address these issues in future studies. For the purpose of aggregated power assessment of entire data center, this performance was considered acceptable with example synthetic signals visualized in Fig. 6. A novel scaling method is proposed and described in the *Algorithm*. The real and synthetic signals are padded and summed to represent transients of extreme (Max) and random (R1 and R2) overlap of stopping time.

The transient swings range from 50% to 12%, and support the need for on-going efforts for managing highly variable training loads. Example mitigation techniques include methods for balancing workload scheduling and reducing the power variation needed by the models during training itself [23, 24].

V. Conclusion

Data center load growth forecasts were summarized to assess their potential future electricity demand in the US. An example energy dispatch was employed to demonstrate coordinated control of distributed energy resources, optimizing the utilization of renewable generation. A review of data center architecture was conducted, including the electric power and cooling infrastructure. Ultra-fast digital twins were proposed as a method for second-level GPU transient analysis and future power infrastructure studies with hybrid energy dispatch. A first attempt to quantify transient swings in data centers employing generative AI shows the importance of future computational workload management development.

Acknowledgment

This paper is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship Grant No. 2239063. The support of PPL Corporation, University of Kentucky the L. Stanley Pigman Chair in Power Endowment, and of the Lighthouse Beacon Foundation are also gratefully acknowledged. Any findings and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the sponsor organizations.

References

[1] J. Aljbour, T. Wilson, and P. Patel, "Powering intelligence: Analyzing artificial intelligence and data center energy consumption," Electric Power Research Institute (EPRI), Palo Alto, CA, Tech. Rep. 3002028905, May 2024, [Online]. Available: https://www.epri.com/research/products/000000003002028905.

[2] J. Noffsinger, M. Patel, and P. Sachdeva. (2025, Apr.) The cost of compute: A $7 trillion race to scale data centers. McKinsey & Company. [Online]. Available: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers.

[3] V. Lee, P. Seshadri, C. O'Niell, A. Choudhary, B. Holstege, and S. A. Deutscher. (2025, Jan.) Breaking barriers to data center growth. Boston Consulting Group (BCG). [Online]. Available: https://www.bcg.com/publications/2025/breaking-barriers-data-center-growth.

[4] B. Chalamala *et al.*, "Data center growth and grid readiness (tr131)," IEEE Power & Energy Society, Tech. Rep., May 2025, [Online]. Available: https://doi.org/10.17023/W4WY-S557.

[5] Y. Li, M. Mughees, Y. Chen, and Y. R. Li, "The unseen ai disruptions for power grids: Llm-induced transients," *arXiv preprint arXiv:2409.11416*, Sep. 2024, [Online]. Available: https://arxiv.org/pdf/2409.11416.

[6] A. Tschand *et al.*, "Mlperf power: Benchmarking the energy efficiency of machine learning systems from watts to mwatts for sustainable ai," in *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025, pp. 1201–1216.

[7] A. Shehabi *et al.*, "2024 united states data center energy usage report," Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA, Tech. Rep. LBNL-2001637, Dec. 2024. [Online]. Available: https://eta.lbl.gov/publications/2024-united-states-data-center-energy-usage-report

[8] International Energy Agency (IEA), "Energy and ai," https://www.iea.org/reports/energy-and-ai, Paris, France, Apr. 2025, licence: CC BY 4.0.

[9] V. Lee, "The impact of genai on electricity: How genai is fueling the data center boom in the u.s." LinkedIn, Boston Consulting Group (BCG), Sep. 2023. [Online]. Available: https://www.linkedin.com/pulse/impact-genai-electricity-how-fueling-data-center-boom-vivian-lee/

[10] National Renewable Energy Laboratory (NREL). (2025) Accelerating speed to power: Data viewer. [Online]. Available: https://maps.nrel.gov/speed-to-power/data-viewer.

[11] M. Turner, "Data center capital of the world: A strategy for a changing paradigm," Loudoun County Board of Supervisors, Loudoun County, VA, Tech. Rep., Oct. 2025, [Online]. Available: https://www.loudoun.gov/ArchiveCenter/ViewFile/Item/13979.

[12] EPRI, "Utility experiences and trends regarding data centers: 2024 survey," EPRI, Palo Alto, CA, Tech. Rep. 3002030643, Sep. 2024, [Online]. [Online]. Available: https://www.epri.com/research/products/000000003002030643

[13] K. A. Kyeremeh, R. E. Alden, A. Patrick, and D. M. Ionel, "Spatiotemporal wind energy assessment for transmission network integration considering the location of electrical substations and loads," in *Proc. 2024 13th Int. Conf. Renewable Energy Res. Appl. (ICRERA)*, 2024, pp. 1805–1810.

[14] M. Abdou-Tankari, G. Lefebvre, J. Arkhangelski, A. Drame, M. Garba, and D. Abdourahimou, "Power quality challenges and urban microgrid-based grid resiliency: Case of Niamey City electrical grid," in *Proc. 2023 12th International Conference on Renewable Energy Research and Applications (ICRERA)*, Oshawa, Canada, 2023.

[15] R. E. Alden, C. Halloran, D. D. Lewis, D. M. Ionel, and M. McCulloch, "Assessment of land and renewable energy resource potential for regional power system integration with ML spatio-temporal clustering," in *Proc. 2023 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, Oshawa, ON, Canada, Aug. 2023, pp. 618–624.

[16] D. D. Lewis *et al.*, "Decarbonization analysis for thermal generation and regionally integrated large-scale renewables based on minutely optimal dispatch with a kentucky case study," *Energies*, vol. 16, no. 4, p. 1999, Feb. 2023, [Online]. Available: https://www.mdpi.com/1996-1073/16/4/1999.

[17] PJM Interconnection. (2025) Data miner — feed directory. Web portal. PJM Interconnection LLC. [Online]. Available: https://dataminer2.pjm.com/list

[18] I. Riepin, T. Brown, and V. M. Zavala, "Spatio-temporal load shifting for truly clean computing," *Advances in Applied Energy*, vol. 17, p. 100202, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666792424000404

[19] N. Wasson, B. Ferreira, and P. Mattson, "Mlperf training benchmark," 2024, [Online]. Available: https://mlcommons.org/benchmarks/training.

[20] A. Rochd, M. Laamim, A. Benazzouz, M. Kissaoui, A. Raihani, and J. M. Guerrero, "Home energy management systems (HEMS) control strategies testing and validation: Design of a laboratory setup for power hardware-in-the-loop (PHIL) considering multi-timescale co-simulation at the smart grids test lab, morocco," in *Proc. 2023 12th International Conference on Renewable Energy Research and Applications (ICRERA)*, Oshawa, Canada, 2023.

[21] S. B. Poore, R. E. Alden, H. Gong, and D. M. Ionel, "Multiphysics and artificial intelligence models for digital twin implementations of residential electric loads," in *2022 11th International Conference on Renewable Energy Research and Application (ICRERA)*, 2022, pp. 576–581.

[22] T. Kaneva, I. Valova, K. Gabrovska-Evstatieva, and B. Evstatiev, "A data-driven approach for generating synthetic load profiles with gans," *Applied Sciences*, vol. 15, no. 14, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/14/7835

[23] Y. Liang, N. Ruan, L. Yi, and X. Su, "An approach to workload generation for modern data centers: A view from alibaba trace," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 4, no. 1, p. 100164, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772485924000164

[24] Z. Wang *et al.*, "Wlb-llm: workload-balanced 4d parallelism for large language model training," in *Proceedings of the 19th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI '25. USA: USENIX Association, 2025.