Short-Term Load Forecasting for AI-Data Center

Mariam Mughees, Yuzhuo Li, Yize Chen, and Yunwei Ryan Li*

* Department of Electrical and Computer Engineering
University of Alberta, Edmonton, Alberta, Canada
{mughees, yuzhuo, yize.chen, yunwei.li}@ualberta.ca

Abstract—Recent research shows large-scale AI-centric data centers could experience rapid fluctuations in power demand due to varying computation loads, such as sudden spikes from inference or interruption of training large language models (LLMs). As a consequence, such huge and fluctuating power demand pose significant challenges to both data center and power utility operation. Accurate short-term power forecasting allows data centers and utilities to dynamically allocate resources and power large computing clusters as required. However, due to the complex data center power usage patterns and the black-box nature of the underlying AI algorithms running in data centers, explicit modeling of AI-data center is quite challenging. Alternatively, to deal with this emerging load forecasting problem, we propose a data-driven workflow to model and predict the short-term electricity load in an AI-data center, and such workflow is compatible with learning-based algorithms such as LSTM, GRU, 1D-CNN. We validate our framework, which achieves decent accuracy on data center GPU short-term power consumption. This provides opportunity for improved power management and sustainable data center operations.

I. INTRODUCTION

The data center industry is expanding rapidly, driven by increasing cloud services, Artificial Intelligence (AI)/Machine Learning (ML) advancements, and data storage needs. Global AI-related electricity demand is projected to grow significantly due to technological, economic, and social factors [1]. This decade has seen major tech companies, like Google and Oracle, invest heavily in new and expanded data center facilities globally [2], [3]. However, these growing demands pose challenges for power grids, leading to concerns about whether they can handle these novel, high-power-density loads, as highlighted in PJM's report on congested lines and rising electricity costs [4]. Additionally, environmental considerations push data centers toward modular designs and renewable energy. AI-focused data centers, with unprecedented per-rack power densities, introduce significant power grid transients, akin to those from Electric Vehicles (EVs) and renewable energy sources [5].

Data center load can introduce huge and instant power variations during data center cold start, shutdown, load shifting or sudden interruption of load. As these large data centers are consuming tens to hundreds of MW power, and may add up to several MW of power changes in few seconds, this can affect grid's frequency control and may demand more responsive frequency regulations in the system. Recently, data centers also operate many power electronics equipment such UPS, control switches and rectifiers. During a transient event, these non-linear components introduce harmonics into the grid and reduce power quality. Thus to improve grid
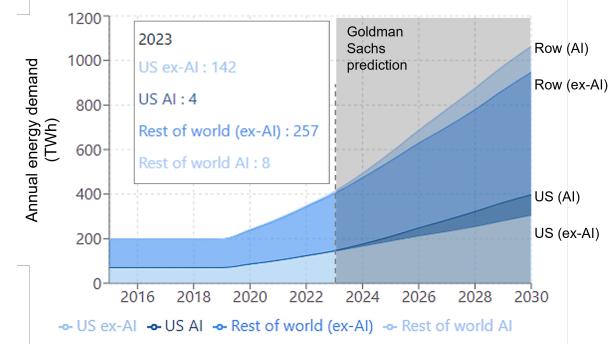

Fig. 1: The stack-area diagram of data center annual energy demand (adapted from [6]). (ex-AI: excluding AI's demand.)

reliability, forecasting the energy consumption is necessary for control strategies such as Automatic Generation Control (AGC), particularly when handling very large power transients caused by AI loads, harmonics and frequency variations [7]. Predicting sags and swells upcoming in power consumption can also help AGC to prepare for these fast ramp-up events.

Existing literature [8]–[10] consider different aspects such as workloads and failure rates, and these features are energy-related. However even diverse datasets with unidentified jobs make predictions difficult to implement and [11] also highlights opportunities like reinforcement learning to improve scheduling. [12], [13] discuss Graphic Processing Unit (GPU) energy management by considering active and idle status according to usage and also based on prediction. [14] considers regression techniques such as Auto Regressive Integrated Moving Average (ARIMA) and fault tree to predict power and failure events in data center facilities. [10] highlights research related to user behavior and how user interaction can affect power. A convolution neural network (CNN) based technique is presented in [15] which considers GPU workloads power consumption, especially by large language models (LLMs) and provides better results than ARIMA. A Deep Neural Network (DNN) is proposed to predict the computational cost of LLM model training in cloud [16]. Short-term data center power forecasting is important specifically for the dynamic and resource-intensive nature of AI and for high-performance computing workloads [17]. Traditional models like ARIMA struggle with complex patterns in high-dimensional data. In contrast, adopting DNNs such as LSTM, GRU, and 1D-CNN can excel in forecasting power consumption for multivariate time series.

This is evident from the literature that data center power consumption is a complex and critical task and depends on many factors such as equipment installed, workload types and operating conditions. As collected data is very versatile in nature and has many sudden dips and peaks which makes forecasting an important and critical task. In this work, we address the challenge of the lack of quality datasets for analyzing energy-intensive GPU-based AI workloads, particularly for training large language models (LLM). We design a general workflow using LSTM, GRU, and 1D-CNN architectures trained on the MIT Supercloud dataset, capturing detailed GPU power metrics over 8 months with a 1-second granularity and a 300- look back window predicts power consumption 90 seconds ahead. Results in the manuscript will show these models achieve decent prediction accuracy, validating their effectiveness for short-term power forecasting in AI-data centers.

II. DATA CENTER LOAD FORECASTING

A. Problem Formulation

The rapid growth of AI computing has transformed data center power requirements. Modern AI workloads feature higher power densities (300W–1,200W per GPU), rapid power fluctuations (e.g., >132 kW/s at the rack level with NVIDIA GB200 NVL72 [18]), and complex, non-linear scaling behaviors. Accurately forecasting these dynamics is crucial for infrastructure design, operational stability, cost optimization, and capacity planning to meet the rising demands of AI workloads. A typical data center is depicted in Fig.2, where computing infrastructure and supporting facilities are main loads, highlighting the complex power infrastructure necessary for an AI-centric data center.

Defining the Time Series Data: Let $X(t)$ represent the load at time step t. The input data for each forecasting point will include the load values from the previous H steps, since the future load is related to the short-term history load recorded. This means that for a given time step t, the input sequence $\mathbf{X}_h$ is defined as:

$$\mathbf{X}_h = \{X(t-H), X(t-H+1), \ldots, X(t-1)\}. \quad (1)$$

Here, $\mathbf{X}_h$ is a vector of load values of H elements, representing the load history up to time $H-1$. The goal is to predict the load values for the next 90 steps based on the input sequence $X(t)$. At time step t, define the forecasting sequence Y_f as:

$$\mathbf{Y}_p = \{Y(t), Y(t+1), \ldots, Y(t+P)\}. \quad (2)$$

In the above formulation, $\mathbf{Y}_p$ denotes a vector of predicted load values over the next P predicted time steps. Let $f(\cdot)$ represent the forecasting function (e.g., a neural network) that maps the input sequence $\mathbf{X}_h$ to the output sequence $\mathbf{Y}_p$:

$$\mathbf{Y}_p = f(\mathbf{X}_h). \quad (3)$$

In real-world systems, load patterns are influenced by many different factors such as environmental factors, user behaviors, and demographics of infrastructure. This makes load forecasting a challenging task. As in this study data-driven forecasting
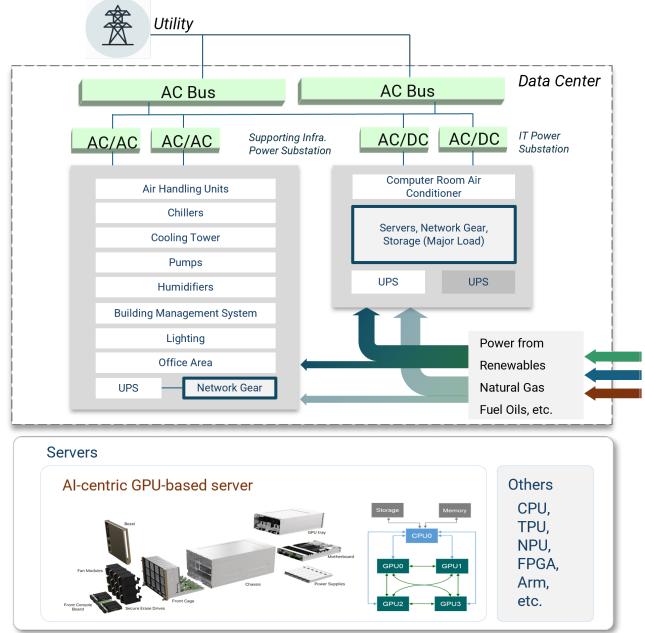

Fig. 2: The electricity demand of AI-centric data center [7].

techniques employed which help model to learn dynamics and nonlinearity from previous data. To fulfill this purpose, the number of look-back and forecasting horizon are defined and utilized.

To train the data center load forecasting model f, we can minimize the mean squared error (MSE) over the forecasting horizon, given by:

$$\text{MSE} = \frac{1}{P} \sum_{i=0}^{P} \left(Y(t+i) - \hat{Y}(t+i) \right)^2; \quad (4)$$

where $\hat{Y}(t+i)$ represents the predicted load at time $t+i$, and $Y(t+i)$ represents the actual load. P is the length of forecasting horizon.

B. Short-Term Data-Center GPU-Power Forecasting

Near to real-time forecasting can help achieve improvements in many folds such as grid, cost reduction and load management. To achieve this forecasting goal, we design and follow a three-stage workflow. Fig.3 outlines our proposed AI-based workflow for short-term GPU load forecasting through time-series prediction algorithms, divided into three main stages:

1) Data Collection and Pre-Processing: The data can be captured through either the hardware or software-based way. Hardware capture can be done through power monitor devices by sensing the power cords of the GPU and CPU motherboard or Power Supply Unit inside the rack. For software capture, the computing unit is based on GPU, CPU or others as mentioned in Fig.2, then, vendor-specific commands can be used to facilitate the collection process. For instance, `nvidia-smi` command can be considered to measure Nvidia's GPU power consumption. Collected data will include GPU power consumption (constitutes the majority
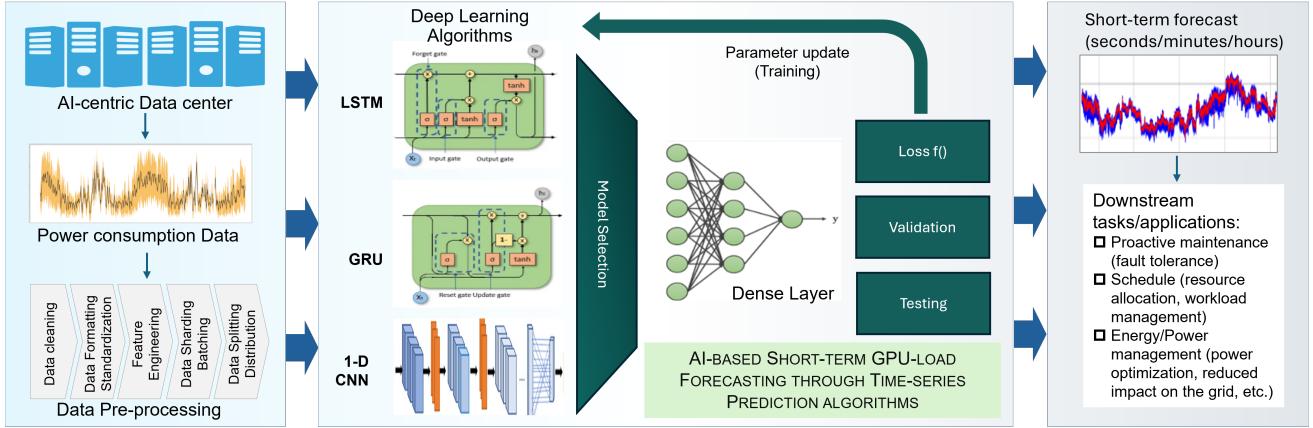
Fig. 3: The workflow of the AI-based short-term forecasting of AI-data center through time-series prediction algorithms

of the total power consumption in AI-centric workloads), memory utilization, GPU temperature, and storage. After the data collection, raw data undergoes several pre-processing steps, including pre-processing, Min-Max normalization, data slicing, etc., to prepare it for feeding into deep learning models.

2) Model Training: Different deep learning architectures, e.g., LSTM, GRU, and 1-D CNN, can be deployed for time-series forecasting. These three models are selected due to their superior capabilities in dealing with sequential data. The model undergoes training, parameter updates, and performance validation to optimize accuracy.

3) Forecasting and Application: The model outputs short-term forecasts of different lengths of periods in advance (e.g., from seconds to minutes) based on the target various down-stream applications, such as proactive maintenance, resource scheduling, and energy/power management.

C. Time-Series Models

There are quite a few AI/ML methods that have been proposed for time-series prediction tasks, while DNNs are getting significant importance in fields of time series forecast-ing as these networks are able to grab complex details from temporal patterns and have multiple layers hidden between input and output. Algorithms like Recurrent neural networks (RNNs) use back propagation through time (BPTT) which helps memorize and analyze information from past time series. RNNs are used for continuous data and are very powerful for capturing dynamics of sequence data. However, these methods suffer from problem of vanishing or exploding gradients when trained on very long data sequences. To handle such issues, idea of an explicit memory augmentation is being implemented in practice in LSTM network. The specifically designed mem-ory cell functions as gated leaky neuron, which has a self-connection to itself at next step and has unity weight, so it duplicates its own value and adds the external signal. And this self-connection is multiplicatively gated by another unit which decides when to clear memory [19], [20].

III. NUMERICAL SIMULATIONS

A. Data Pre-processing on Real-world Data Center GPU-Load Dataset

In this study, we address the data center power consumption forecasting problem using a real-world dataset from the MIT Supercloud [21], a high-performance computing (HPC) system (GPU: Nvidia Volta V100, CPU: Intel Xeon Gold 6248). The dataset spans February to October 2021 and includes 100-millisecond interval logs of GPU/CPU utilization, scheduling details, and physical parameters like temperature. Key GPU metrics include power, memory, utilization, and temperature, with anonymized user data organized by job ID and node. Ag-gregated GPU power consumption peaks at 45 kW across 448 GPUs. The dataset details workload composition, dominated by vision networks (e.g., U-Net: 1,431 jobs; VGG, ResNet, and Inception follow), language models (e.g., BERT: 189 jobs; DistillBERT: 172 jobs), and graph neural networks (e.g., SchNet, DimeNet: lower counts). Pre-processing maintains a 1-second granularity, with power consumption aggregated by job ID and node to reflect total power drawn from the local distribution system. After normalization via a min-max scaler, the data uses a 300-lookback window to predict 90 seconds ahead. Fig.4 illustrates the GPU power consumption trends and train-validation-test splits (ratios: 0.7, 0.15, 0.15).

B. Simulation Results

Prediction results are compared on different metrics in Table I such as RMSE, MAE, sMAPE, and R-squared. As mentioned above, LSTM, GRU and 1D-CNN are being considered for prediction methods. Fully connected LSTM has consistently achieved the best results with the lowest RMSE, MAE, sMAPE and highest R-squared value which is an indicator of robust prediction and low error. GRUs show slightly declined performance in comparison to LSTM, in terms of RMSE and R-squared error and MDB indicates slight negative bias ness in prediction results. While 1D-CNN has a relatively lower per-formance and noticeably more positive bias. 1-minute zoomed graphs show predictions at a finer granularity. The model's
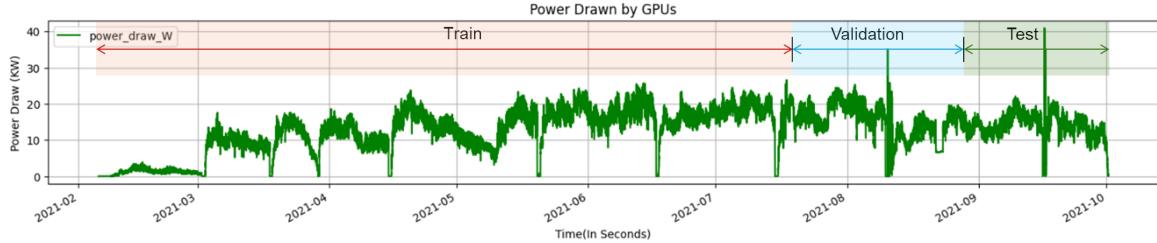
Fig. 4: MIT Supercloud Dataset used in the simulation and the data split.
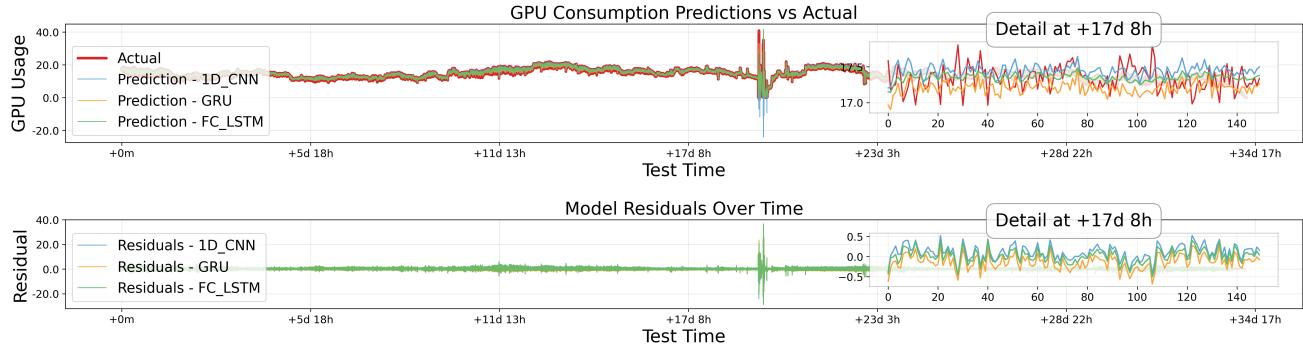

Fig. 5: Prediction results (upper) and prediction error in terms of residuals (lower) for 1D_CNN, GRU, and LSTM. Zoom-in view is visualized to the right.
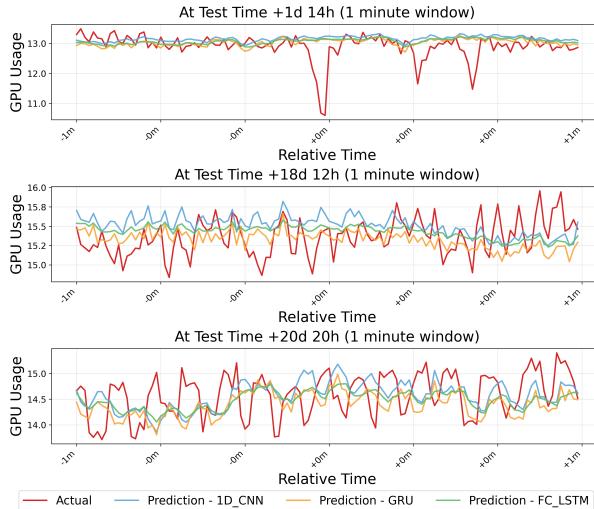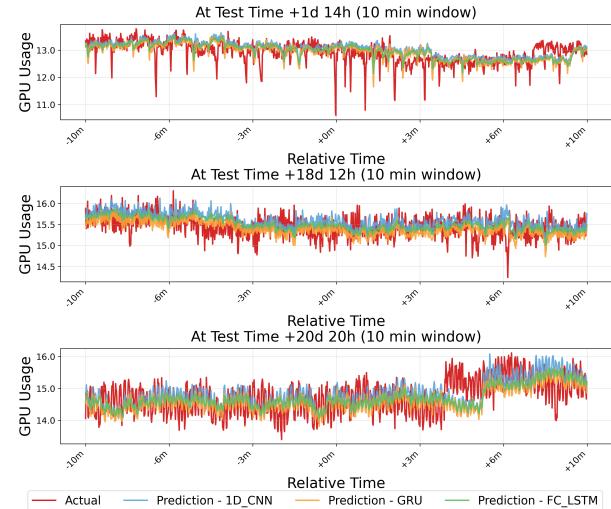

Fig. 6: Prediction results (1 minute).

Fig. 7: Prediction results (10 minutes).

ability to capture short-term fluctuations and variations can be observed from these graphs. 1-minute forecasting is presented in Fig.6, results closely aligning with actual values. However, for sudden dips and peaks, prediction struggles to capture pattern. 10-minute zoomed plots display predictions over a longer time frame in Fig.7, as these graphs present a broader trend of prediction. Forecasting shows a consistent trend with real data and noticeable deviations can be seen around spikes where prediction lags the actual value. Fig.8 shows zoomed-in predictions vs. actual GPU consumption over a one-hour interval, depicting close tracking but with some deviations, especially in high-variability periods.

C. Discussions and Recommendations

This data center short-term forecasting can be considered sufficient for grid response as power generation has two main layers of control: primary control and secondary control. The primary control is the load frequency control which is dependent on inertia and governor operation and the time to response is 2-10 seconds. Secondary control is AGC which also maintains frequency and power balance in the longer
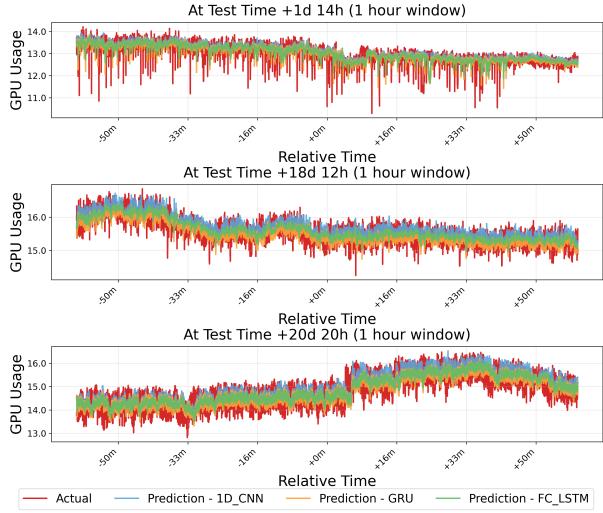
Fig. 8: Prediction results (1 hour).

TABLE I: Performance metrics for different models

Model	RMSE	MAE	MBD	sMAPE	R_squared
FC_LSTM	0.5624	0.2916	0.0319	2.2557	0.9639
GRU	0.5668	0.3163	-0.0433	2.6288	0.9634
1D_CNN	0.5789	0.3172	0.1216	2.5014	0.9618

RMSE (Root Mean Square Error): Measures the average magnitude of the errors, indicating overall accuracy.
MAE (Mean Absolute Error): Represents the average absolute difference between predicted and actual values.
MBD (Mean Bias Deviation): Measures the average bias or tendency of predictions, with positive/negative values indicating over/underestimation.
sMAPE (Symmetric Mean Absolute Percentage Error): A normalized measure of accuracy expressed as a percentage.
R_squared (Coefficient of Determination): Indicates the proportion of the variance in the dependent variable explained by the model.

term with a response time of 10-30 seconds. While considering load-side management, this prediction analysis can provide an added advantage in the event of peak shaving with the support of battery energy storage systems (BESSs). Batteries with a capacity of several MW can be deployed in an AI-extensive system and can help in load shifting, energy arbitrage and grid resiliency. For instance, batteries can be charged during times of low electricity costs and discharged when costs are high, reducing operational expenses and huge demand spikes for data centers. This strategy can be beneficial in markets with variable electricity prices, such as those driven by real-time pricing or time-of-use rates.

IV. CONCLUSIONS

In this paper, a short term forecasting technique is implemented on fine-grained GPU power dataset. Workload in this data-driven approach contains training of LLMs and various AI algorithms, which highlights very dynamic power consumption nature of data centers in the wild. As data centers increasingly adopt AI-intensive jobs, accurate power prediction becomes more critical. Future research in this direction will consider more impact factors of power consumption and include more robust algorithms like liquid neural networks for resilient and energy-efficient data center operations.

REFERENCES

[1] L. Lin and A. A. Chien, "Adapting datacenter capacity for greener datacenters and grid," in *Proceedings of the 14th ACM International Conference on Future Energy Systems*, pp. 200–213, 2023.

[2] C. Staff, "Google pours billions into new u.s. data centers: Here's where," 2024. Accessed: 2024-11-16.

[3] D. Staff, "Oracle's larry ellison: We're building out 100 data centers globally," 2024. Accessed: 2024-11-16.

[4] A. E. Research, "Data center load growth in pjm." https://auroraer.com/insight/data-center-load-growth-in-pjm/, 2024. [Accessed 11-17-2024].

[5] L. Lin, R. Wijayawardana, V. Rao, H. Nguyen, E. W. GNIBGA, and A. A. Chien, "Exploding ai power use: an opportunity to rethink grid planning and management," in *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, pp. 434–441, 2024.

[6] G. Sachs, "Ai is poised to drive 160% increase in data center power demand." Goldman Sachs Insights, May 2024. Accessed: 2024-11-18.

[7] 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*, 2024.

[8] G. Amvrosiadis, J. W. Park, G. R. Ganger, G. A. Gibson, E. Baseman, and N. DeBardeleben, "On the diversity of cluster workloads and its impact on research results," in *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, pp. 533–546, 2018.

[9] M. Blöcher, L. Wang, P. Eugster, and M. Schmidt, "Switches for hire: Resource scheduling for data center in-network computing," in *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 268–285, 2021.

[10] G. Wilkins, S. Keshav, and R. Mortier, "Hybrid heterogeneous clusters can lower the energy consumption of llm inference workloads," in *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, e-Energy '24, (New York, NY, USA), p. 506–513, Association for Computing Machinery, 2024.

[11] S. Wang, S. Chen, and Y. Shi, "Utilization-prediction-aware energy optimization approach for heterogeneous gpu clusters," *The Journal of Supercomputing*, vol. 80, pp. 9554–9578, May 2024.

[12] M. Rossi and D. Brunelli, "Forecasting data centers power consumption with the holt-winters method," in *2015 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS) Proceedings*, pp. 210–214, IEEE, 2015.

[13] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power," *ACM SIGARCH Computer Architecture News*, vol. 37, no. 1, pp. 205–216, 2009.

[14] H. Shoukourian and D. Kranzlmüller, "Forecasting power-efficiency related key performance indicators for modern data centers using lstms," *Future Generation Computer Systems*, vol. 112, pp. 362–382, 2020.

[15] L. Bai, W. Ji, Q. Li, X. Yao, W. Xin, and W. Zhu, "Dnnabacus: Toward accurate computational cost prediction for deep neural networks," 2022.

[16] P. Patel, E. Choukse, C. Zhang, I. n. Goiri, B. Warrier, N. Mahalingam, and R. Bianchini, "Characterizing power management opportunities for llms in the cloud," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS '24, (New York, NY, USA), p. 207–222, Association for Computing Machinery, 2024.

[17] Q. Hu, P. Sun, S. Yan, Y. Wen, and T. Zhang, "Characterization and prediction of deep learning workloads in large-scale gpu datacenters," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[18] Schneider Electric and NVIDIA, "Ai reference designs to enable adoption: A collaboration between schneider electric and nvidia," white paper, Schneider Electric.

[19] Z. Ye, W. Gao, Q. Hu, P. Sun, X. Wang, Y. Luo, T. Zhang, and Y. Wen, "Deep learning workload scheduling in gpu datacenters: A survey," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–38, 2024.

[20] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in optimizing recurrent networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8624–8628, IEEE, 2013.

[21] S. Samsi, M. L. Weiss, D. Bestor, B. Li, M. Jones, A. Reuther, D. Edelman, W. Arcand, C. Byun, J. Holodnack, *et al.*, "The mit supercloud dataset," in *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–8, IEEE, 2021.