Voltage Regulation in Distribution Systems with Data Center Loads

Yize Chen*, Baosen Zhang†
* Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada
†Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, USA
Emails: yize.chen@ualberta.ca, zhangbao@uw.edu

***Abstract*—The recent boom in foundation models and AI computing has raised growing concerns about the power and energy demand of large-scale data centers. This paper focuses on the voltage issues caused by the volatility of data center power demand, which also aligns with recent observations of more frequent voltage disturbances in power grids. To address these challenges, we propose a dynamic voltage control scheme by harnessing data center's load regulation capabilities. By taking local voltage measurements and adjusting power injections at each data center bus through the dynamic voltage and frequency scaling (DVFS) scheme, we are able to maintain safe voltage magnitude in a distributed fashion. Simulations using real large language model (LLM) inference loads validate the effectiveness of our proposed mechanism.**

***Index Terms*—Grid strength, inverter-based resource, power flow, power system stability, system strength**

I. Introduction

Large-scale data centers are in the midst of a period of explosive growth driven by advances in AI and compute. The International Energy Agency (IEA) estimated that data centers consumed 460 TWh in 2022, accounting for 2% of global electricity consumption–a figure that is still rapidly increasing and is expected to reach 5% by 2030 [1]. Such energy burdens are generally observed in both the training and deployment stages of modern AI models [2], and are further exacerbated by the growing cooling needs for more advanced GPU clusters [3], [4].

Much attention has recently been paid to both the average and peak power demand of data centers. As electric power represents more than 60% of the total operational cost of data centers [5], they face the complex task of meeting customer quality of service requirements while balancing cost and sustainability. Meanwhile, system operators and utility companies are having a difficult time coping with the ever-increasing and fluctuating data center demand. Consequently, a number of works have proposed to reduce the total or peak power demand of data centers [2], [6].

In this paper, we do not directly address either of these issues. Instead, we focus on voltage regulations in grids with data centers, a problem we believe to be critical to their integration into the power system but has perhaps received less attention. Data centers, as well as many other devices, are sensitive to the voltage magnitudes of their grid connections. A short voltage disturbance in Northern Virginia in July 2024 led to more than 60 hyperscale data centers disconnecting from the grid and almost caused widespread blackout in the PJM system [7]. A review by NERC also highlighted the challenges with voltage in data center load management [8].

Voltage regulation itself is a classical problem in power system operations. Most distribution systems have several types of equipment dedicated to voltage support, including tap-changing transformers, switched capacitors and STATCOMs. In systems with inverter-based resources, many algorithms have been proposed to use the power electronics in the inverters to inject reactive power for voltage support (see, e.g. [9]–[11] and the references therein).

Fig. 1: Power consumption curve of a GPU during the model training of GPT2 124M. Multiple GPUs are synchronized during training, creating volatile and large disturbances on the grid.

What makes data center grid integration challenging and interesting in the context of voltage regulation is their unique load characteristics especially in the era of LLMs and foundational models [12]. Fig. 1 shows the single GPU's power consumption profiles of training a GPT2 124M model on a cluster consisting of Nvidia A100 GPUs. Sec. II-B shows some additional consumption profiles of performing several inference tasks. We observe sharp dips in power consumption during both inference and training.[1] In these dips, a large down ramp is quickly followed (sometimes within a second) by a large up ramp, which can lead to large voltage deviations in the distribution system if left uncompensated. Because of the suddenness of the load changes, it is difficult to compensate for them using conventional devices, which cannot act quickly enough to deal with (sub)second-level transients [9]. In addition, large data centers are often placed at the end of large lines, and since reactive power injections tend to

[1]These dips are much more important in transient dynamics compared to more "steady-state" type of analysis. They are too short to have much impact on total energy consumption, and because they represent a decrease in power draw, they would not factor into peak demand.

979-8-3315-8138-1

be "localized", power electronic devices placed away from the data centers would have limited impact, despite their fast reaction capabilities.

Therefore, in this paper, we turn data centers from voltage disrupters to regulators, and analyze how data centers could proactively adjust their active and reactive power injections to regulate voltage. For reactive power compensation, we adopt the standard technique that utilizes power electronic rectifiers to modulate reactive power injections. To change the active power, we design a novel dynamic voltage and frequency scaling (DVFS) scheme to control the dynamic power draw of the GPUs. More specifically, we focus mostly on changing the frequency of the GPUs [13]. For example, NVIDIA GA100 supports operations at different frequencies in the range between 210 MHz and 1410 MHz [14], while modern operating systems and hardware are capable of changing voltage and frequency states at intervals ranging from several milliseconds up to 100 milliseconds with latency up to 100ms, making it a realistic control knob for voltage regulation tasks. Using DVFS to adjust the frequency and power consumption of the underlying GPU, it is possible to achieve optimization at the iteration level (spanning only milliseconds) compared to the overall inference times (spanning several seconds). As the power draw scales linearly with switching frequencies [15], we can trade off compute with power. DVFS is a common technique for circuit design to manage chip-level power [16], but has not been leveraged to support grid-level operations.

Because many distribution systems do not yet have real-time communication capabilities, the controllers in this paper have the flavor of primary controllers, where they take local measurements and adjust power injections at their own buses. We present the models in Section II. We show how DVFS can be adapted to form an active power control loop (Section III) that is quite effective in regulating voltage as shown through simulations in Section IV.

II. Data Centers in the Grid

A. Distribution Grid Voltage Regulation

We consider a power network with $N+1$ buses, numbered as $0, 1, \ldots, N$, with bus 0 as the feeder. We assume that the network is a tree with the feeder at the root. We write $i \to j$ if j is a child bus of j. Let $\boldsymbol{v}$ be the vector of voltage magnitudes where V_i is the voltage magnitude of bus i. Given a branch (i, j) in the tree, let r_{ij} and x_{ij} denote the resistance and impedance of the line, respectively. We use $P_{i,j}$ and $Q_{i,j}$ to denote the active and reactive power flowing from i to j, respectively. We use P_i and Q_i to denote power injections at bus i, and use vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ to denote injections at all buses. Voltage and power are related by DistFlow equation [17], [18]:

$$\sum_{k,j\to k} P_{jk} = P_{ij} - r_{ij}\frac{P_{ij}^2 + Q_{ij}^2}{V_i^2} + P_j \quad \forall j$$

$$\sum_{k,j\to k} Q_{ij} = Q_{ij} - x_{ij}\frac{P_{ij}^2 + Q_{ij}^2}{V_i^2} + Q_j \quad \forall j$$

$$V_j^2 - V_k^2 = 2(r_{jk}P_{jk} + x_{jk}Q_{jk}) - (r_{jk}^2 + x_{jk}^2)\frac{P_{jk}^2 + Q_{jk}^2}{V_j^2} \quad \forall (j,k)$$

We view the above equations as defining an implicit function from $\boldsymbol{p}$ and $\boldsymbol{q}$ to $\boldsymbol{v}$, written as $\boldsymbol{v} = f(\boldsymbol{p}, \boldsymbol{Q})$. Several techniques exist to efficiently compute this function when the network is a tree [19], [20]. We use the nonlinear DistFlow equations since we are interested in large loads that undergo rapid changes and capturing the nonlinear effect between voltage and power injections becomes important.

A standard requirement for distribution networks is that voltages should not deviate too far from their rated value, for example, not more than 5% [21]. As we show, just using standard equipment, such as tap-changing transformers and switched capacitors, is not effective when large and fast-changing data center loads are present.

In this paper, we assume some buses are data centers, some buses have inverter-based resources (e.g., solar PV or battery storage) and the rest of the buses are PQ loads. Our goal is to use the flexibility in data centers and inverter-based resources to control the voltages to be close to the rated voltage, while minimizing the cost of using these resources. Since the control of inverter-based resources is not the focus of this paper, we assume that they run one of the numerous algorithms (for example, droop control on reactive power [22]) that have been proposed and would otherwise not deal with their behavior until the simulation results.

B. Load Profile of Common Inference Tasks

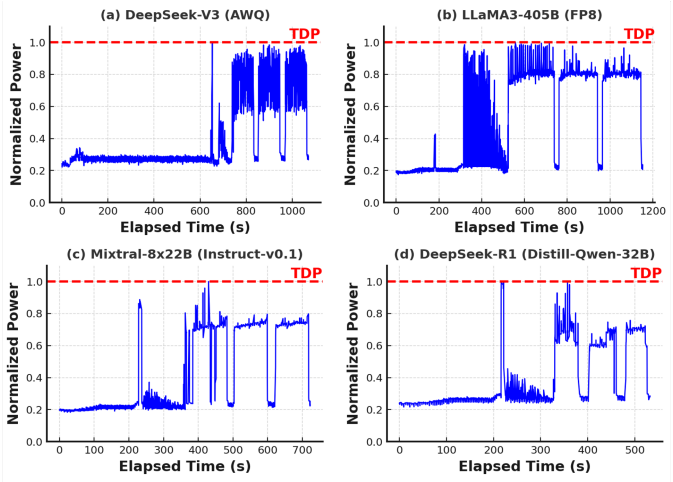

Fig. 2: Power profiles of the four LLM inference tasks reveal distinct power ramping characteristics. Power is normalized in $[0, 1]$ range with thermal design power (TDP) limit marked.

In order to model data center load and characterize associated flexibility, we collect real measurements of power consumption profile by performing inference tasks using several common foundational models, and these inference timeseries is collected on an 8×A100 cluster instance as shown in each subfigure in Fig. 2. Such load exhibits high ramps, which is synchronized across multiple GPUs. Also note that GPU power can be slightly over the Thermal design power (TDP) when GPU frequency is higher.

C. Control of AI Data Center Loads

We now turn to the control of data center loads and observe that the standard method of controlling reactive power to regulate voltage may not be a good strategy. First, most large data centers already employ power factor correction and operate at close to unity power factor. Second, data centers also operate close to rated capacity, which can constrain reactive power capability of the power electronic converters even if they deviate from unity power factor [23], [24]. Third, because of the high r/x ratios of power lines in distribution systems, voltage is sometimes more sensitive to changes in active power. Therefore, we consider two axes control on both active and reactive power.

Since we do not assume the availability of real-time communication, $\boldsymbol{p}$ and $\boldsymbol{q}$ need to be updated based on local voltage measurements. We consider a discrete time system indexed by time t. To serve AI load, the desired active power consumption of the data center at bus i and time t is denoted as $P^{\text{ref}}_{i,t}$. We assume that $P^{\text{ref}}_{i,t+1}$ is predictable at time t, that is, we have complete knowledge of the data center load in the next time step.[2] After observing its voltage $V_{i,t}$ at time t, bus i sets its active power consumption in the next time step as $P_{i,t+1} = u^P_{i,t}(V_{i,t}, P^{\text{ref}}_{i,t})$. Similarly, the reactive power at time $t+1$ is $Q_{i,t+1} = u^Q_{i,t}(V_{i,t})$, except that there is no desired value for the reactive power. The system dynamics becomes

$$\boldsymbol{v}_{t+1} = f(\boldsymbol{u}^P_t, \boldsymbol{u}^Q_t). \tag{2}$$

We are interested in finding u^P_i and u^Q_i for a controllable bus i that regulate the voltage close to its reference value V^{ref}_i, satisfy their respective constraints, and do not severely degrade the quality of service (QoS) of the data centers. We examine the problem over a total horizon of length T, with T being a large number. We require that the active and reactive powers are bounded, that is, $u^P_i \in [\underline{P}_i, \overline{P}_i]$ and $u^Q_i \in [\underline{Q}_i, \overline{Q}_i]$.[3] To ensure the compute quality of service, we assume that the total power (or energy) consumed remains invariant, that is, $\sum_t P_{i,t} = \sum_t P^{\text{ref}}_{i,t}$ for a bus i with a data center. Together, the controller functions u^P_i and u^Q_i should be designed such that the following constraints are satisfied:

$$P_{i,t+1} = u^P_{i,t}(V_{i,t}, P^{\text{ref}}_{i,t}), \;\; Q_{i,t+1} = u^Q_{i,t}(V_{i,t}) \tag{3a}$$

$$u^P_i \in [\underline{P}_i, \overline{P}_i], \;\; u^Q_i \in [\underline{Q}_i, \overline{Q}_i] \tag{3b}$$

$$\boldsymbol{v}_{t+1} = f(\boldsymbol{u}^P_t, \boldsymbol{u}^Q_t) \tag{3c}$$

$$\sum_t P_{i,t} = \sum_t P^{\text{ref}}_{i,t}\text{, } i \text{ is a data center bus.} \tag{3d}$$

Of course, our goal is to minimize the total deviation from the reference voltage $\sum_{t=1}^T |V_{i,t} - V^{\text{ref}}_i|$, but designing the optimal controller is nontrivial. In the rest of paper, we present a simple yet effective strategy that performs well in experiments, and leave the question of optimal design to a future paper.

[2]This is a reasonable assumption since the compute load of applications in data centers is predictable over a short timescale [25].

[3]A constraint on the apparent power can also be used.

III. Data Center Load Control

Here we adopt a simple strategy based on a modified linear droop control. The control knob we tune to control active power is the DVFS capability on the GPU chips. We assume that the GPUs within a data center are running with a common clock, which is the case when a single compute-intensive AI application is being served. Let $f_{\text{clk},i}$ denote the clock frequency and let $\tilde{P}_i$ denote the power consumed at a data center without any other constraints or controls.

It turns out that GPU power is approximately linearly related to frequency: $\tilde{P}_i \sim f_{\text{clk},i}$ [26].The constant of proportionality can depend on a number of factors, including the exact hardware setup and the compute application. In turn, we control and set the clock frequency as a linear function of the voltage deviation $f_{\text{clk},i} \sim (V_{i,t} - V^{\text{ref}}_i)$. Together, we obtain an active power control droop in the form of $k_{p,i}(V_{i,t} - 1)$, with $\overline{P}_i$ and $\underline{P}_i$ determined by GPU i's highest and lowest frequency available. Such a controller is fast and distributed, which makes it practical for data center control tasks.

Because of the total power balance constraint (3d), we cannot just set u^P_i based on the standard voltage/power droop. Therefore, we introduce a term $S_{i,t}$ that keeps track of the power not served up to time t with $S_{i,0} := 0$ which is also translated to non-served AI load. And the active power control on bus i is designed as

$$u^P_{i+1,t} = [P^{\text{ref}}_{i,t} - k_{p,i}(V_{i,t} - 1) + \alpha_{i,t} S_{i,t}]^{\overline{P}_i}_{\underline{P}_i}, \tag{4}$$

where $[a]^{\overline{a}}_{\underline{a}}$ thresholds a variable to the upper or lower bound ($[a]^{\overline{a}}_{\underline{a}} = \max(\min(a, \overline{a}), \underline{a})$). And $S_{i,t}$ evolves as the running sum of unserved power demand for bus i. We update it as

$$S_{i,t+1} = S_{i,t} + P^{\text{ref}}_{i,t} - u^P_{i+1,t}. \tag{5}$$

The $\alpha_{i,t}$ parameter is tuned to encourage the satisfaction of the total power balance constraint. For example, if $S_{i,t}$ is large (compute load not served), then $\alpha_{i,t}$ would be increased.

The reactive power control loop takes the standard form of

$$u^Q_{i,t+1} = [u^Q_{i,t} - k_{q,i}(V_{i,t} - 1)]^{\overline{Q}_i}_{\underline{Q}_i}. \tag{6}$$

We assume that data centers have limited capacity for reactive control, and set $\overline{Q}_i$ and $\underline{Q}_i$ to be smaller than their active power counterparts. For buses with inverters but not data centers, we assume that they only perform reactive power control.

Proving the convergence of a system controlled by (4) and (6) is an important future direction for us. On one hand, (thresholded) linear droop controllers are ubiquitous in power systems. On the other hand, we are only aware of stability results for linearized systems (e.g., ones using LinDistFlow models [27]). Nevertheless, the controllers converges quickly and performs well in all of our simulations.

IV. Numerical Simulations

A. Simulation Setup

Data Center Load: The load presented in this study (e.g., Fig. 2) utilizes 8 A100 GPUs offering 8×80GB of high-bandwidth memory, allowing simultaneous data-parallel and

model-parallel strategies for LLM inference tasks. We use standard set of prompts and query the LLama 3.1 70B Instruct Model. In practice, we see high GPU utilization when the GPUs are loaded with LLM inference tasks. Since neural transformer-based LLM models are autoregressive which generate output on a token-by-token basis, they require multiple iterations and the inference duration are diverse and depending on the query length. Correspondingly we observe large and fast swings in power draw, particularly during steps involving heavy matrix multiplications and data communications. When the LLM query firstly arrives or inference task finishes, there are large ramping up or ramping down in the power demand.

After collecting the power measurement curves under a variety of prompts, we scale the GPU load to mimic the load behavior of a small- to medium-sized data center with peak workloads up to $280\,kW$. We then randomly select snippets of such data center loads and assign them randomly to the data center nodes, which simulates the random arrival patterns for LLM tasks in AI data centers. We use the standard first-come, first-serve (FCFS) scheduling policy for LLM queries.

Power Grid Simulations: We modify the standard IEEE 123-bus by randomly selecting nodes to equip with inverters or data centers. For the inverter buses, we assign the same bounds for reactive power injections, that is, the inverters are sized uniformly. We keep the number of inverters and inverter buses fixed in the network, but assign 1 to 5 random data center buses to examine the effects of high penetration of data center loads. An extreme case with 8 data centers are also tested. We perturb the other loads by adding Gaussian noise. The active power is balanced by slack bus's active power injections at every timestep t. We also quantify LLM serving delay based on job completion time after implementing DVFS. Throughout the simulations, we set the simulation interval and control interval as 1 second. We note that the proposed droop-based control strategy is computationally very light and easy to implement.

B. Simulation Results

It can be revealed in Fig. 1 and Fig. 2 that to accommodate large number of LLM weights and matrix operations, both LLM training and inference tasks exhibit large power ramps. Due to the nature of autoregressive model, the valley-peak ramping behaviors exhibit longer periodic patterns, while the query length and thus peak load time is highly diverse. During inference stages, larger LLMs such as Llama3 405B do not always exhibit more dynamic power behavior—mid-scale models like Mixtral-8x22B and DeepSeek-V3-AWQ, which show more pronounced fluctuations. The 10–20% ramping range dominates, indicating frequent computations tied to attention mechanisms inside LLMs.

Fig. 3 visualizes the voltage deviation from the nominal values under three settings: plain data center integrations without inverter nor data center load control. It can be clearly observed that without inverter or data center voltage control, the added data center load will heavily affect the voltage profile, and a large voltage deviation is observed. With growing number of data centers from 1 to 8, average voltage deviation tends

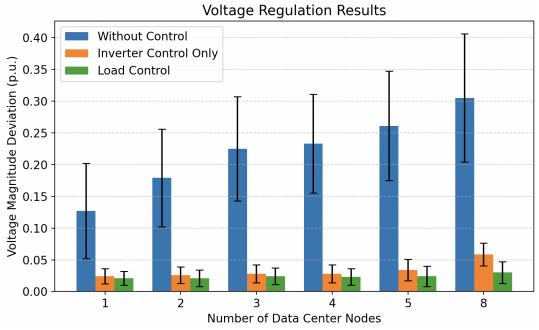

Fig. 3: Voltage magnitude deviation (p.u.) and standard deviation with varying number of AI data center load buses.

to increase for all three settings. While it is noteworthy by integrating data center load control, the least deviation in voltage profile is recorded. More than 97% of the buses always have a voltage deviation within $0.05\,p.u.$ throughout the simulation horizon. On average, such a control scheme achieves more than 12.8% gain compared to inverter-based droop control only in terms of voltage magnitude deviation.

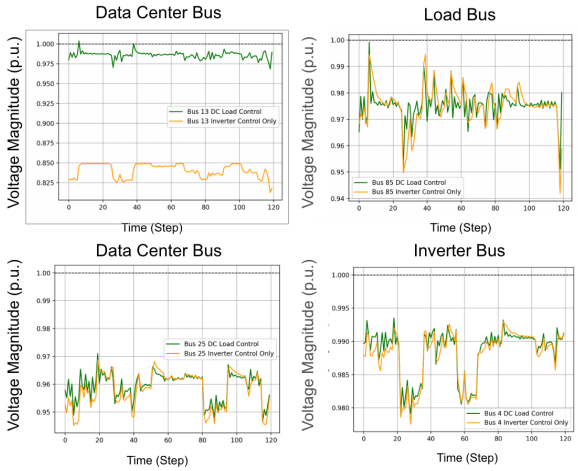

Fig. 4: 2-minute voltage curves for selected two data center buses, a load bus, and an inverter bus. Control with only inverters (orange) and also with data centers (green) are compared.

We evaluate the voltage regulation effects with varying droop parameters k_p. The results are summarized in Table. I. For all choices of k_p and varying degree of data center integrations (number of data centers from 1 to 5), the buses' average voltage deviation can be controlled within $0.03\,p.u.$. There is a mild increase in voltage deviation when the number of data center grows, as it becomes more challenging to regulate more volatile load. With $k_p = 10$ and with 5 data center load buses, proposed coordinated control of data centers and inverters can achieve an average voltage magnitude deviation of 0.024, while with only inverter control, such deviation can reach 0.034. Moreover, as shown in Fig. 4, bus 13's voltage is out of the $[0.95, 1.05]$ p.u. range with inverter-only control. This bus has a large data center load and 4 neighbors, making it difficult to control. Further, as illustrated in Fig. 4, with data center actively participating the voltage regulation, voltage magnitude for all buses are closer to 1.0 p.u..

TABLE I
VOLTAGE MAGNITUDE DEVIATION AND LLM QUERY DELAY WITH VARYING NUMBER OF DATA CENTER NODES IN 123-BUS SYSTEM

No. of Data Centers	$k_p = 1$		$k_p = 10$		$k_p = 20$	
	ΔV (p.u.)	Average Delay (s)	ΔV (p.u.)	Average Delay (s)	ΔV (p.u.)	Average Delay (s)
1	0.020 ± 0.012	0.3	0.020 ± 0.013	6.0	0.019 ± 0.012	11.2
2	0.021 ± 0.012	0.5	0.021 ± 0.013	5.6	0.020 ± 0.011	9.2
3	0.026 ± 0.013	0.5	0.024 ± 0.013	6.2	0.022 ± 0.013	13.7
4	0.027 ± 0.015	1.1	0.023 ± 0.013	6.0	0.021 ± 0.015	13.8
5	0.028 ± 0.017	1.4	0.024 ± 0.016	6.0	0.024 ± 0.016	17.5

For each data center, the k_p parameter also decides the tradeoff between load flexibility and voltage regulation participation. As indicated in Table. I, when $k_p = 20$, the LLM inference query experiences a significant delay compared to smaller k_p. This is because when there is a ramping up in active load curve, the droop control strategy implicitly reduce such fast load increases, and shift them to latter intervals. Similarly, original LLM tasks are having sudden endings when the end-of-generation tokens are generated. While under droop control, the data center gradually decreases load. Thus, in practice, striking a balance in the choice of the k_p parameter is essential for optimizing the performance of data centers under dynamic load and voltage conditions.

V. CONCLUSION

In this work, we look into the voltage issues with growing data centers connecting to the grid. To effectively regulate distribution grid voltage, we show it is possible to convert data center as flexible load resources by utilizing GPU's DVFS capabilities. By leveraging a distributed droop control scheme, we propose a simple and effective linear control strategy that dynamically adjust LLM workload as well as inverter buses' injections. Simulation results on real LLM inference load validate the effectiveness of our proposed approach, demonstrating significant improvements in grid voltage profiles while balancing LLM query delays. In future work, we will integrate more principled AI load forecasting model to aid the controller, and investigate more fine-grained control knobs and explore the integration of these computing-side control mechanisms with existing grid-side management systems.

REFERENCES

[1] International Energy Agency, "Electricity 2024," https://www.iea.org/reports/electricity-2024, 2024.

[2] A. A. Chien, L. Lin, H. Nguyen, V. Rao, T. Sharma, and R. Wijayawardana, "Reducing the carbon impact of generative ai inference (today and in 2035)," in *Proceedings of the 2nd workshop on sustainable computer systems*, 2023, pp. 1–7.

[3] S. Nagendra, "Thermal analysis for nvidia gtx480 fermi gpu architecture," *arXiv preprint arXiv:2403.16239*, 2024.

[4] Y. Zhu, J. Zou, and R. C. Pilawa-Podgurski, "A 1500-a/48-v-to-1-v switching bus converter for next-generation ultra-high-power processors," *IEEE Transactions on Power Electronics*, 2024.

[5] J. J. Jensen, "Data centers and energy: Reaching sustainability, the vital role of heat reuse," *Danfoss*, 2024, accessed: 2025-04-30.

[6] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening geographical load balancing," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 1, pp. 193–204, 2011.

[7] M. Gooding, "Virginia narrowly avoided power cuts when 60 data centers dropped off the grid at once," *Data Center Dynamics*, 2025.

[8] NERC, "Incident review considering simultaneous voltage-sensitive load reductions," 2025, accessed: 2025-04-30.

[9] H.-G. Yeh, D. F. Gayme, and S. H. Low, "Adaptive var control for distribution circuits with photovoltaic generators," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1656–1663, 2012.

[10] B. Zhang, A. Y. Lam, A. D. Domínguez-García, and D. Tse, "An optimal and distributed method for voltage regulation in power distribution systems," *IEEE Transactions on Power Systems*, vol. 30, no. 4, 2014.

[11] P. Srivastava, R. Haider, V. J. Nair, V. Venkataramanan, A. M. Annaswamy, and A. K. Srivastava, "Voltage regulation in distribution grids: A survey," *Annual Reviews in Control*, vol. 55, pp. 165–181, 2023.

[12] 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.

[13] W. Bao, C. Hong, S. Chunduri, S. Krishnamoorthy, L.-N. Pouchet, F. Rastello, and P. Sadayappan, "Static and dynamic frequency scaling on multicore cpus," *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 13, no. 4, pp. 1–26, 2016.

[14] W. Luo, R. Fan, Z. Li, D. Du, Q. Wang, and X. Chu, "Benchmarking and dissecting the nvidia hopper gpu architecture," in *2024 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2024, pp. 656–667.

[15] A. Agarwal and J. Lang, *Foundations of analog and digital electronic circuits*. Elsevier, 2005.

[16] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital integrated circuits*. Prentice hall Englewood Cliffs, 2002, vol. 2.

[17] M. Baran and F. F. Wu, "Optimal sizing of capacitors placed on a radial distribution system," *IEEE Transactions on power Delivery*, vol. 4, no. 1, pp. 735–743, 2002.

[18] M. Farivar and S. H. Low, "Branch flow model: Relaxations and convexification—part i," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2554–2564, 2013.

[19] H.-D. Chiang and M. E. Baran, "On the existence and uniqueness of load flow solution for radial distribution power networks," *IEEE Transactions on Circuits and Systems*, vol. 37, no. 3, pp. 410–416, 2002.

[20] B. Fang, C. Zhao, and S. H. Low, "Convergence of backward/forward sweep for power flow solution in radial networks," *IEEE Transactions on Control of Network Systems*, 2025.

[21] N. E. M. Association *et al.*, *American National Standard for Electric Power Systems and Equipment-Voltage Ratings (60 Hertz)*. National Electrical Manufacturers Association, 1996.

[22] W. Cui, J. Li, and B. Zhang, "Decentralized safe reinforcement learning for inverter-based voltage control," *Electric Power Systems Research*, vol. 211, p. 108609, 2022.

[23] K. Turitsyn, P. Sulc, S. Backhaus, and M. Chertkov, "Options for control of reactive power by distributed photovoltaic generators," *Proceedings of the IEEE*, vol. 99, no. 6, pp. 1063–1073, 2011.

[24] L. Streitmatter, T. Joswig-Jones, and B. Zhang, "Geometry of the feasible output regions of grid-interfacing inverters with current limits," *IEEE Control Systems Letters*, 2025.

[25] M. Mughees, Y. Li, Y. Chen, and Y. R. Li, "Short-term load forecasting for ai-data center," in *2025 IEEE Power Energy Society General Meeting (PESGM)*, 2025, pp. 1–5.

[26] J. Yu, A. Taneja, J. Lin, and M. Zhang, "Voltanallm: Feedback-driven frequency control and state-space routing for energy-efficient llm serving," 2025. [Online]. Available: https://arxiv.org/abs/2509.04827

[27] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: Optimality and stability analysis," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3794–3803, 2015.