Journal of Contemporary Science and Technology Management
Volume 02, Issue 04, December 2026

The Role of Artificial Intelligence in Optimizing Data Center Energy Consumption and Electrical Grid Performance

Anthony Defede
Application Engineer, Armistead Mechanical, Inc., New Jersey, United States

Abstract

Artificial intelligence (AI) is both a major new source of electricity demand and one of the most promising tools for managing it. Data centers consumed about 415 TWh of electricity in 2024, and the International Energy Agency (IEA, 2025) projects this will more than double to around 945 TWh by 2030, driven largely by AI workloads. This article examines how AI itself can reduce that burden. It reviews three areas of application: AI-based control of cooling and facility systems, intelligent scheduling and resource allocation of computing workloads, and AI tools that improve the performance of the electrical grid. Evidence from deployed systems shows that machine learning can manage cooling at the scale of a full data center floor, shift flexible computing to cleaner and less congested hours, and help grids carry more power on existing lines. The article also discusses limits, including data quality, trust in automated control, and the risk that efficiency gains are outpaced by demand growth. It concludes with recommendations for operators, utilities, and policymakers.

Keywords: Artificial intelligence; data centers; energy efficiency; cooling optimization; workload scheduling; grid performance

1. Introduction

AI presents an energy paradox. Training and running large AI models requires dense clusters of graphics processing units (GPUs) that draw far more power than conventional servers. Hyperscale AI campuses now commonly exceed 100 MW, and some are planned at gigawatt scale (Chen et al., 2025). Anioke (2025) shows that this growth raises total electricity demand, creates new risks for grid stability, and complicates the transition to sustainable energy.

Yet the same technology offers ways to manage these effects. Data centers and power grids are complex systems with thousands of interacting variables: temperatures, fan speeds, workloads, power flows, prices, and weather. Human operators and fixed control rules cannot optimize all of these together in real time. Machine learning models can. They can learn how a system behaves from its own data, predict what will happen next, and choose actions that save energy while keeping operations within safe limits.

This article reviews how AI is being applied to optimize data center energy use and grid performance. It draws on Anioke's (2025, 2026) work on grid impacts and stranded power, Lakavath's (2023) analysis of serverless cloud computing, and studies of AI-based cooling, carbon-aware scheduling, and grid optimization.

2. The Energy Challenge Created by AI Workloads

The IEA (2025) estimates that data centers used about 1.5% of global electricity in 2024, with consumption growing around 12% per year since 2017. Electricity use in AI-focused data centers is projected to more than quadruple by 2030. In the United States, Shehabi et al. (2024) report that data

center electricity use rose from 58 TWh in 2014 to 176 TWh in 2023, about 4.4% of national consumption. They project 6.7% to 12.0% by 2028.

Inside a data center, energy goes to two main uses. The first is the IT equipment itself: servers, storage, and networking. The second is the supporting infrastructure, especially cooling, which removes the heat that IT equipment produces. Both offer room for optimization. IT energy depends on how efficiently computing work is packed onto hardware. Cooling energy depends on how well the facility matches its cooling effort to the actual heat load.

AI workloads also behave differently on the grid. Chen et al. (2025) note that their power demand can swing sharply within seconds as training jobs start and stop, that they connect through power electronic converters, and that they cluster in a few regions. These features make AI data centers harder to integrate. They also make them good candidates for intelligent control, because a load that changes quickly can, in principle, be managed quickly.

3. AI for Cooling and Facility Energy Management

Cooling is the largest energy use in a data center after the IT equipment, and it is where AI control has its strongest record. Traditional cooling systems run on fixed rules and conservative set points, set by engineers to avoid overheating under worst-case conditions. This leaves much of the cooling effort unused for long periods.

Machine learning offers a more precise approach. Lazic et al. (2018), working with Google's data center team, developed a model-predictive control system that learns how a data center floor responds to changes in fans, water flow, and other cooling settings. The model was trained on operating data, then used to choose control actions that kept temperatures within limits while using less energy. The researchers showed that the learned controller could regulate conditions on a real server floor, using only a few hours of exploratory data, and improve efficiency relative to the existing controllers.

The principle extends beyond cooling. AI models can forecast heat loads from scheduled computing work, detect faulty equipment from unusual sensor patterns, and coordinate cooling with on-site batteries and generators. Each of these reduces wasted energy and the power a facility must draw from the grid at peak times.

4. AI for Workload Scheduling and Resource Optimization

The second area of opportunity lies in how computing work is assigned to hardware and to time.

Better use of servers. Idle and underused servers still consume significant power. Cloud platforms address this by sharing hardware across many users and scaling resources up and down with demand. Lakavath (2023) examines serverless computing on Microsoft Azure Functions, a model in which applications run only when triggered and resources scale automatically. Lakavath (2023) assesses its performance, scalability, and cost efficiency. Because users pay only for the computing they actually use, and the platform can scale resources down when demand falls, serverless designs encourage higher utilization of shared hardware. Higher utilization means less idle power per unit of useful work. AI-driven autoscaling, which predicts demand and allocates resources ahead of time, makes this approach more efficient still.

Carbon-aware and grid-aware scheduling. Many computing tasks do not need to run at a precise moment. Model training, batch analytics, and data processing can often wait a few hours. Radovanović et al. (2023) describe Google's carbon-intelligent computing system, which uses machine learning to

forecast both next-day grid carbon intensity and the computing demand of each data center cluster. It then sets hourly limits on available capacity, shifting flexible tasks toward hours when cleaner power is available. The system operates across Google's global fleet and also lowers peak power demand at each site, which helps the local grid.

Using stranded power. Anioke (2026) describes stranded power as energy or capacity that exists but goes unused, such as renewable output curtailed because of transmission limits, or reserve capacity inside a facility that is rarely drawn. AI scheduling can direct flexible workloads toward times and places where such surplus power is available, turning wasted energy into useful computing. The same forecasting tools that predict carbon intensity can predict when curtailment is likely.

5. AI for Electrical Grid Performance

AI also improves the grid that serves data centers and every other customer.

Forecasting. Accurate forecasts of demand and of wind and solar output are the foundation of reliable grid operation. Machine learning models can combine weather data, historical load, and real-time measurements to predict both more accurately than traditional methods. Better forecasts reduce the reserves operators must hold and make it easier to integrate variable renewable energy.

Getting more from existing lines. Transmission lines are usually rated conservatively, assuming worst-case weather. In practice, cooler temperatures or wind allow lines to carry more power safely. AI-based tools can estimate real-time line capacity from weather and sensor data. The IEA (2025) estimates that up to 175 GW of additional transmission capacity could be unlocked in existing lines through wider use of AI. This matters for data centers, because grid connection delays are now one of the main barriers to new projects.

Fault detection and maintenance. AI can detect early signs of equipment failure from sensor data and spot unusual events on the network faster than human operators. This supports faster restoration after faults and more targeted maintenance. Chen et al. (2025) stress that better monitoring and modeling of data center behavior are also needed, since these loads can respond to grid disturbances in unexpected ways.

Coordinating data centers with the grid. The greatest gains come when AI on both sides works together. A grid operator's forecasts can be shared with a data center's scheduling system, so that flexible workloads move away from hours of grid stress. Anioke (2025) argues that this kind of coordination is central to managing data center growth sustainably.

6. Challenges and Limitations

AI optimization is not a complete solution, and several limits deserve attention.

First, AI tools depend on good data. Sensors must be accurate and well maintained, and models must be retrained as equipment and workloads change. Poor data can produce poor decisions.

Second, operators must trust automated control. A data center's first priority is uptime. Systems such as the one described by Lazic et al. (2018) therefore operate within strict safety limits and allow human override. Similar safeguards are needed for any AI that acts on grid equipment.

Third, AI consumes energy itself. The models used for optimization are usually small compared with large language models, but their energy use should still be counted.

Fourth, efficiency gains may be outpaced by demand growth. The IEA (2025) shows that even with strong efficiency improvements, data center electricity use is expected to rise substantially. AI optimization can slow the growth and shift it in time, but it cannot by itself remove the need for new generation and transmission.

Finally, flexibility has costs. AI data centers are very capital-intensive, so operators are reluctant to delay work often (IEA, 2025). Scheduling gains are greatest for workloads that can genuinely wait.

7. Recommendations

Five steps would help realize the benefits of AI optimization:

1. **Adopt AI-based facility control** for cooling and power systems, with clear safety limits and human oversight (Lazic et al., 2018).
2. **Expand carbon-aware and grid-aware scheduling** of flexible workloads, using forecasts of grid conditions (Radovanović et al., 2023).
3. **Raise hardware utilization** through cloud and serverless designs that scale resources to actual demand (Lakavath, 2023).
4. **Deploy AI tools on the grid** for forecasting, real-time line ratings, and fault detection (IEA, 2025).
5. **Create data-sharing and market frameworks** that let grid operators and data centers coordinate, and that reward data centers for using surplus or stranded power (Anioke, 2026).

8. Conclusion

AI is driving a rapid rise in data center electricity demand, but it also provides some of the most effective tools for managing that demand. Machine learning can control cooling systems with a precision that fixed rules cannot match. Intelligent scheduling can raise server utilization, shift flexible work to cleaner hours, and make use of stranded power. On the grid, AI can improve forecasts, unlock capacity on existing lines, and detect faults earlier. None of these measures removes the need for new infrastructure, and each depends on good data and careful safeguards. Together, however, they can make data centers more efficient, grids more capable, and the relationship between the two more cooperative. The challenge for operators, utilities, and policymakers is to adopt these tools quickly enough to keep pace with the growth they are meant to manage.

References

Anioke, C. (2025). How data centers affect the electrical grid: Implications for electricity demand, grid stability, and energy sustainability. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 17*(2), 61–66.

Anioke, C. (2026). Stranded power from data centers and possible uses. *Journal of Emerging Intelligence and Engineering Technologies, 3*(1), 41–49.

Chen, X., Wang, X., Colacelli, A., Lee, M., & Xie, L. (2025). *Electricity demand and grid impacts of AI data centers: Challenges and prospects* (arXiv:2509.07218). arXiv. https://doi.org/10.48550/arXiv.2509.07218

International Energy Agency. (2025). *Energy and AI*. IEA. https://www.iea.org/reports/energy-and-ai

Lakavath, S. R. (2023). Serverless computing on Azure Functions: Performance, scalability, and cost efficiency. *International Journal of Technology, Management and Humanities, 9*(4), 377–387. https://doi.org/10.21590/ijtmh.9.4.420

Lazic, N., Lu, T., Boutilier, C., Ryu, M. K., Wong, E. J., Roy, B., & Imwalle, G. (2018). Data center cooling using model-predictive control. In *Advances in Neural Information Processing Systems 31* (pp. 3818–3827). Curran Associates.

Radovanović, A., Koningstein, R., Schneider, I., Chen, B., Duarte, A., Roy, B., Xiao, D., Haridasan, M., Hung, P., Care, N., Talukdar, S., Mullen, E., Smith, K., Cottman, M., & Cirne, W. (2023). Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems, 38*(2), 1270–1280. https://doi.org/10.1109/TPWRS.2022.3173250

Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A. B., Holecek, B., Koomey, J., Masanet, E., & Sartor, D. (2024). *2024 United States data center energy usage report* (LBNL-2001637). Lawrence Berkeley National Laboratory. https://doi.org/10.71468/P1WC7Q