International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Green Data Center: Reduced Energy Consumption During Peak Load

Ayush Nautiyal

Scholar, Computer Engineering and Applications, Magalayatan University

Abstract

Earlier, Data Centre sustainability has defined as an efficiency problem that can be resolved by building better hardware etc., improving airflow, increasing reliance on renewable energy, and lower the facility overheads. This definition is now changed; the current wave of AI-centric growth is creating a peak load problem more than the annual energy problem. In 2024, data-centres consumed around 415 TWh of electricity which is approximately 1.5% of the global demand, and it is expected to increase to 945 TWh by 2030. [1] As per the estimates by International Energy Agency (IEA), 60% of electricity is consumed by hardware infrastructure in modern data centre, while cooling consumption ranged from 7% in efficient hyperscale data-centre to more than 30% in traditional data-centre. [1]

Peak load plays very significant role because data-centre are designed to handle the highest demand intervals. Requirements to serve the peak load is considered for the requirement to grid connection, transformer, UPS systems, cooling systems etc. even if the average load is lower.

This paper formulates technical framework for reduction of energy consumption during the peak load in green data centre. It assesses dynamic voltage and frequency scaling (DVFS), virtualization, cooling, workload scheduling, demand response, power capping, thermal storage, renewable energy, grid-interactive UPS and high-density thermal architecture. Primary finding suggests that the peak load reduction needs to be referred as portfolio problem: the maximum benefit arise when adaptable IT controls are integrated with facility side thermal optimization and supply side flexibility. The case study referred in the paper support this finding. Google has reported the deployment of demand response for ML workload; AWS reported that mechanical energy consumption reduced by 46% during peak cooling by new designs; Microsoft reported system level avoidance of potential CO_2 through grid interactive UPS programme in Ireland; OVHcloud reported reduction of cooling electricity by 50% and water use by 30% through new AI-enabled cooling architecture. (Google, 2025; AWS, 2025; Microsoft, 2022; OVHcloud, 2025). [2]

The contribution of this paper is to set of engineering models for peak shaving, demand-response potential, and emission effects. The paper priorities primary, official, and recent peer reviewed sources. The key conclusion of this paper is that peak load reduction need to be designed across all integrated layers of data centre – IT, facility, and supply layers.

Keywords: Green data centre, peak load reduction, power usage effectiveness, demand response, DVFS, AI cooling optimization, Peak Shaving

1. Introduction

The research problem addressed in this paper is: how green data centre can reduce the energy consumption during peak load without impacting the reliability and service quality. This problem is very critical as in the current context data centres are not just minor consumer of electricity. As per IEA estimates data centre electricity demand increased to 415 TWH in 2024 that is a growth of 12% per year since 2017, and is estimated to be doubled to 945 TWh by 2030. [1]

Earlier, data centre sustainability was framed mainly in terms of Power Usage Effectiveness (PUE) and annual renewable procurement; which is not valid anymore. Introduction of AI and other new technologies rack densities and electrical concentration got increased, and requires quick, automated controls. Uptime's 2024 survey documents accelerating rack power growth, flat average PUE, and growing operational concern about cot and dense IT. [3]

This paper starts with a simple idea: handling peak electricity demand is a tougher and more important challenge than just reducing the overall yearly energy usage. A data centre might cut its total energy consumption, but if it still relies heavily on power during times when the grid is under stress, it can remain expensive, and harder to scale. On the other hand, a data centre that can adjust when and how it uses power, reduce energy spent outside of core operations, and shift its remaining demand to leaner interval of the day has a more impact. It helps emission and supports a more stable grid. That is why this paper combine energy efficiency and interaction, procurement practices, and system reliability; because these are all connected and need to be managed together, not in isolation. [4]

2. Why peak load matters

The technical reason peak load matters is simple: facility power is the sum of IT power plus cooling, power-chain and miscellaneous overhead. The IEA notes that cooling-system share varies from about 7% in efficient hyperscale sites to over 30% in less-efficient enterprise facilities, which immediately explains why cooling optimisation is often the fastest route to peak relief. At the same time, AI workloads raise rack density and power volatility, making load-shaping and forecasting far more valuable than they were in older, steadier enterprise estates.[1]

The economic reason is equally strong. Even when total annual energy remains manageable, short periods of peak use drive tariffs, capacity procurement, utility interconnection timelines and infrastructure spend. Google's 2026 demand-response milestone explicitly frames flexible demand as a way to avoid building transmission and generation designed only for brief peaks, and DOE workshop participants argued that speed-to-connection incentives can be more powerful than conventional demand-response payments. That is a critical practical insight: in fast-growing markets, flexibility is increasingly valuable not only for operating cost, but for getting capacity online sooner.[5]

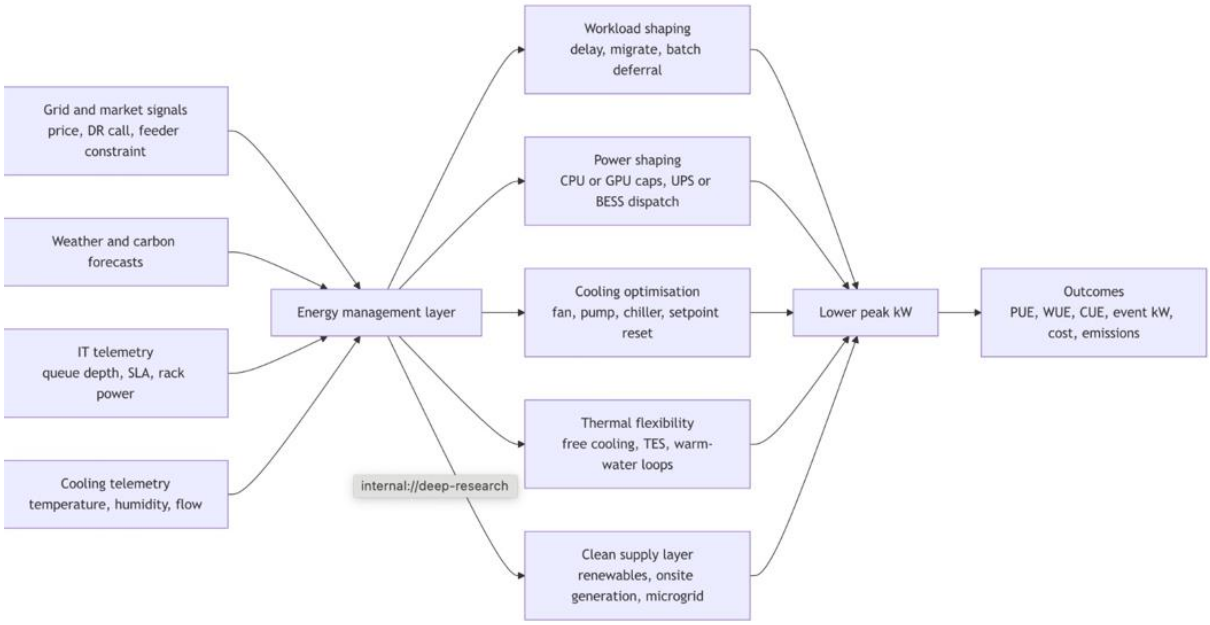
The policy context is also tightening. In Europe, Delegated Regulation (EU) 2024/1364 requires reporting data centres to submit specified information and key performance indicators to the European database on data centres. In the United States, FERC Order No. 2222 is intended to enable aggregated distributed energy resources—explicitly including storage, distributed generation, demand response, energy efficiency and thermal storage—to participate more fully in regional wholesale markets. For data-centre

operators, this means load flexibility is becoming visible both upstream, in regulation, and downstream, in tariffs and market participation designs. [6]

Metrics must also evolve. PUE remains the most common infrastructure metric and is formally defined as facility energy divided by IT energy. Microsoft summarises the same ratio on its datacentre sustainability page, and The Green Grid remains the canonical industry source. But liquid-cooling studies show PUE can flatten or even understate real gains when a measure reduces both infrastructure and IT fan power. That is why practitioners increasingly pair PUE with WUE, CUE, energy reuse metrics, peak kW, and event-based demand-response KPIs. In other words, PUE is necessary, but not sufficient for peak-load management. [7]

3. Technologies and strategies

The current best practice is a hierarchical control stack: software first, thermodynamics second, infrastructure third. Figure 1 is a synthesis of the control layers repeatedly described in DOE workshop findings and in operator case studies from Google, Meta, NTT and CenterSquare.[3]



Demand response and workload shifting. Flexible computing is the cleanest peak-shaving resource because it reduces demand without new hardware. Google's operational demand-response system temporarily limits non-urgent tasks during grid events and reschedules or reroutes them to other times or locations, while preserving user-facing services. In Europe, Google scheduled daily peak-period reductions from 5 pm to 9 pm across multiple countries during winter 2022–23, and by March 2026 it had signed 1 GW of demand-response capacity with U.S. utility partners. In its peer-reviewed carbon-aware work, Google also showed cluster-level virtual capacity curves that cut flexible load by roughly 50% during peak-carbon hours and produced about an 8% power drop over multi-hour periods. [8]

Dynamic provisioning and power capping. Dynamic provisioning reduces active equipment count or compute intensity when demand, temperatures or grid constraints tighten. In practice, this includes right-

sizing active servers, throttling CPUs/GPUs, and linking queue schedulers to power envelopes. DOE's 2025 data-centre summit slides explicitly note "cap processor power during peak times" as an emerging tactic, while recent research on datacentre power capping highlights its role in sustainability-aware optimisation and power-demand shifting. The key caveat is service risk: power capping is most effective when paired with workload classification, so batch and deferrable tasks absorb most of the adjustment while latency-sensitive transactions remain protected. [9]

Cooling optimisation. For existing air-cooled sites, airflow optimisation, chilled-water reset, better containment, and dynamic sequencing of chillers/pumps remain the fastest "green" interventions. CenterSquare's Mesa site is the most transparent recent example: airflow optimisation plus chilled-water optimisation, executed through a dynamic cooling management system using AI/ML, delivered major annual savings and reduced non-IT load while freeing trapped electrical capacity for more IT deployment. NTT's Rhine-Ruhr 1 pilot used a digital twin plus AI to optimise chiller and pump control, delivering 19.1% lower chiller energy in the first months and projecting up to 25% annual savings. These examples matter because they avoid the long lead times of structural retrofits. [10]

Liquid cooling and warm-water operation. For AI/HPC deployments, liquid cooling is increasingly less an option than a necessity. NREL's ESIF demonstrates the architectural end-state: warm-water direct liquid cooling, chiller-less operation, and heat reuse. The Vertiv-NVIDIA study shows the retrofit path: a high-density air-liquid hybrid facility cut total power 10.2% and facility power 18.1% when liquid cooling covered about 75% of the load. The newer HAWK study adds an important control insight: once liquid cooling is deployed, warmer supply temperatures can sharply cut cooling-system power, particularly when outdoor conditions support efficient rejection or reuse. [11]

Free cooling and economisation. The NetApp Bangalore case remains a particularly useful en-IN example because it proves that "free" cooling is not only a Nordic strategy. The facility was designed for 20% full free cooling and 78% partial free cooling, with lowest monthly PUE around 1.35 when fully cooled by outside air. Berkeley Lab and its partners present this as evidence that moderate Indian climates can still support meaningful operational cost and energy savings through outside-air economisation, provided air management and controls are designed carefully. [12]

Thermal storage, renewable integration, on-site generation and energy storage. DOE workshop participants repeatedly identified thermal storage, large-scale batteries, on-site generation and improved controls as central to grid-responsive data centres. NREL's Cold UTES project is the most notable recent innovation: it proposes storing "cold" underground using off-peak electricity and discharging that thermal reserve during peak hours, effectively turning cooling infrastructure into long-duration storage. While still pre-commercial, it directly targets one of the hardest green-data-centre problems: summer cooling peaks that stress both the site and the grid. At the policy interface, FERC Order 2222 and related market reforms increase the strategic value of batteries, thermal storage and distributed generation because they can increasingly be monetised as flexible resources rather than sitting idle as pure backup. [13]

AI/ML for prediction and control. AI is most useful where systems are nonlinear, highly coupled, and too dynamic for static rules—exactly the condition of real data-centre cooling plants and queue-management systems. Google's DeepMind cooling system, Meta's RL airflow control, and NTT's hybrid AI-plus-physics control stack all reached this conclusion independently. The lesson is not "add AI everywhere"; it

is “use AI where there is enough telemetry, enough actuator freedom, and a safe fallback mode.” Meta and NTT both emphasise fallback controls and operational guardrails, which is crucial for conference-level technical credibility. [14]

Table 1: the strategies most directly relevant to peak load.

Strategy	Peak-load mechanism	Best fit	Main caveat	Evidence base
Workload shifting and demand response	Move or defer flexible jobs out of grid-stress hours; optionally reroute geographically	Hyperscale, cloud, batch-heavy estates	Requires scheduler maturity and SLA classification	Google pilots and 1 GW DR milestone; Google carbon-aware paper
Dynamic provisioning and power capping	Reduce active compute or cap CPU/GPU power during events	AI/HPC and mixed estates with controllable jobs	Can affect performance if applied bluntly	DOE summit guidance; recent power-capping research
Airflow and chilled-water optimization	Lower fan, pump and chiller power while maintaining thermal compliance	Existing air-cooled colocation and enterprise sites	Sensor quality and commissioning matter	CenterSquare; NTT Rhine-Ruhr 1
AI/ML cooling control	Predict optimal setpoints under changing weather and IT loads	Sites with rich telemetry and safe override logic	Needs fallback and M&V discipline	DeepMind; Meta RL; NTT digital twin
Liquid cooling and warm-water operation	Shrinks cooling overhead and enables higher supply temperatures	AI, GPU, HPC, high rack density	Retrofit complexity, leak/failure management	NREL ESIF; Vertiv-NVIDIA; HAWK study
Free cooling and economization	Uses ambient conditions to displace mechanical cooling	Cool, dry, or moderate climates; well-managed air paths	Climate dependence; filtration and humidity control	NetApp Bangalore; Meta Luleå operational design note
Thermal	Serves residual	Campuses with	Upfront cost and	DOE workshop

Strategy	Peak-load mechanism	Best fit	Main caveat	Evidence base
Workload shifting and demand response	Move or defer flexible jobs out of grid-stress hours; optionally reroute geographically	Hyperscale, cloud, batch-heavy estates	Requires scheduler maturity and SLA classification	Google pilots and 1 GW DR milestone; Google carbon-aware paper
storage, batteries, and on-site generation	peaks without drawing it all from the grid in real time	tariff exposure or grid constraints	market participation complexity	summary; NREL Cold UTES; FERC 2222 context
Renewable integration and heat reuse	Lowers operating emissions and improves whole-system efficiency	New builds, campuses, colder climates, district-energy contexts	Does not automatically solve instantaneous peak unless paired with storage or flexibility	NetApp Bangalore; NREL ESIF; Meta Luleå

4. Case study evidence

The evidence is stronger when the cases are compared side by side. Some measures clearly reduce peak power demand, while others focus on cutting background or overhead energy use. In practice, this lower overhead often helps reduce how large or how frequent peak loads are, particularly during warmer periods. Where precise peak demand figures are not publicly available, this report describes the peak impact as inferred rather than directly stated.

Operator or study	Location	Period	Intervention	Quantitative result	Cost and timeline signals
Google DeepMind	Google live data-centre site, site not disclosed	2016 deployment	ML cooling optimisation	Cooling energy -40%; overall PUE overhead -15%; lowest site PUE observed	Cost not disclosed; operational rollout followed live testing [5]
Google carbon-aware computing	Google fleet, selected clusters	Paper 2021, journal 2023	Day-ahead virtual capacity curves for flexible workloads	Flexible load drop of about 50% in peak-carbon hours; cluster power about -8% over 3–6 h windows	No public capex; software-led control with global scheduling implications[15]
Google demand response	Europe, Taiwan, Oregon/Nebraska/Southeast U.S.; contracts in U.S.	Pilots 2022–23; milestone 2026	Grid-event load reduction and workload rerouting	Daily peak-period reductions in Europe (5 pm–9 pm); by 2026 1 GW of DR capacity signed with utilities	Cost not disclosed; positioned as a faster alternative to some grid upgrades [16]

CenterSquare Mesa	Mesa/Phoenix, Arizona, U.S.	Case study published 2024	Airflow optimisation plus chilled-water optimisation with AI/ML cooling controls	141.59 kW peak-demand savings; 1,240,369 kWh/year savings; about USD 107,044/year saved	Project cost about USD 150,000 for the cited measure; USD 76,177 rebate; 0.7-year simple payback after rebate [17]
NTT Rhine-Ruhr 1	Bonn, Germany	Pilot from 2025 disclosure	AI/digital-twin optimisation of refrigeration controls	Cooling-system energy use nearly -20%; chillers -19.1% in first months; up to 25% annual savings expected	Hosted on-site; emphasis on fallback control and staged integration [10]
Meta RL pilot	One Meta region, site not disclosed	Pilot started 2021; published 2024	Simulator-based RL for supply-airflow control	Supply-fan energy -20% on average; water use -4%	Cost not disclosed; Meta states method is being rolled out to existing and future sites [12]

NetApp Bangalore	Bengaluru, India	Campus operating since 2017; 2020 data in case study	Outside-air economiser, DRUPS, containment, integrated BMS, renewables	Designed for 20% full and 78% partial free cooling; annual average PUE 1.42; lowest monthly PUE 1.35; annual energy 11,953 MWh; renewables supply >75% of annual electricity; rooftop solar 116 kW	Utility contract demand 4,000 kVA; operational IT load 0.96 MW vs design 4.26 MW [13]
NREL ESIF HPC	Golden, Colorado, U.S.	Operational since opening; 2026 page update	Warm-water direct liquid cooling and waste-heat reuse	Annualised PUE 1.036; waste heat reused as primary heating source; Better Buildings cites estimated annual average ERE 0.7	Showcase facility; chiller-less design [18]

Vertiv-NVIDIA ASME study	Baltimore, Maryland, U.S.	Published 2023	Progressive direct-to-chip liquid cooling retrofit	At 74.9% liquid-cooled load: facility power -18.1%, total data-centre power -10.2%, server-fan power -80%, PUE 1.38 → 1.34	Mid-size 1–2 MW Tier 2 facility; public cost not disclosed [19]
HAWK Applied Energy study	Stuttgart, Germany	Published 2026, using HAWK operational data	Supply-water temperature optimisation in direct liquid cooling	Raising supply water 17°C → 25°C cut liquid-cooling-system power 63.3% at 19°C outdoor wet-bulb temperature	Research study; points to dynamic control potential rather than public retrofit economics [20]

The two patterns stand out from these cases. First, the best near term returns come from software and controls applied to existing plants: CenterSquare, NTT, Meta and DeepMind all show this. Second, the best long term structural performance appears in sites that redesign heat removal and reuse – NREL ESIF, NetApp Bangalore's economiser native design, and liquid cooled AI/HPC studies. That combination strongly suggests a phased roadmap: optimise first, retrofit second, redesign third. [9]

5. Conclusion and actionable recommendations

The evidence does not support a single "silver bullet" for green peak-load reduction. Instead, it supports a portfolio architecture with three progressively stronger layers.

The first layer is software-defined flexibility: queue-aware workload shifting, demand-response integration, predictive cooling control, and selective power capping. This layer is attractive because it produces measurable savings with the smallest construction burden. Real-world examples include DeepMind's 40% cooling-energy reduction, Meta's 20% fan-energy reduction, NTT's 19.1% early chiller

savings, and CenterSquare's 141.59 kW peak reduction with a sub-one-year post-rebate payback. These results make a strong conference-paper finding: the cheapest "green megawatt" is often the megawatt not drawn in the first place. [5]

The second layer is thermal architecture: economisers, containment, better sequencing, liquid cooling, and warmer-water operation. This layer matters more as rack densities rise. The strongest finding here is that liquid cooling is not simply a way to keep hardware safe; it is a way to reposition the whole site on the energy curve. NREL's 1.036 PUE, the Vertiv-NVIDIA 10.2% total-power reduction, and the HAWK study's 63.3% cooling-system-power reduction under warmer-water operation all point in the same direction. For AI-era facilities, cooling design is now peak-load design. [18]

The third layer is grid interactivity: batteries, thermal storage, on-site generation, renewable integration and market participation. This is the least mature layer operationally, but probably the most important strategically. Google's 1 GW demand-response milestone and DOE's workshop conclusions show that data centres are becoming part of power-system planning, not merely passive loads. Where interconnection queues are long, behind-the-meter flexibility may become as important as annual electricity price.

6. Actionable recommendations for operators and researchers

Instrument the site before optimising it. Minimum viable observability should include rack and row power, cooling-loop temperatures and flows, chiller and pump power, weather, queue depth, and SLA class. Without this, AI/ML control becomes guesswork rather than engineering. [12]

Target non-IT overhead first. Existing air-cooled facilities should prioritise airflow optimisation, containment, chilled-water reset, and dynamic fan/chiller control, because these deliver fast savings with modest operational risk. CenterSquare and NTT provide the clearest recent evidence. [9]

Classify workloads by flexibility. Tag jobs as latency-critical, deferrable, migratable, or interruptible; then expose those classes to the energy-management layer. Google's results show that flexible-load shaping is practical at scale when schedulers understand which work can move. [6]

Treat liquid cooling as a systems decision, not a rack decision. The largest gains emerge when liquid cooling is coupled with higher supply-water temperatures, better heat rejection, possible heat reuse, and revised control logic. [19]

Design for event response, not only annual averages. Track event-based KPIs such as peak kW avoided, curtailed MWh, recovery time, and cost avoided—not just annual PUE. Peak-aware metrics align better with utility, policy and interconnection realities. [22]

Use storage selectively. Batteries are valuable for fast electrical peaks; thermal storage is especially attractive where cooling drives the site peak. DOE workshop participants were explicit that thermal storage may be an especially promising alternative in the flexibility stack.[22]

Align engineering with policy and tariffs. Data-centre flexibility has practical value only when contracts, tariffs and market rules recognise it. EU reporting rules and U.S. DER participation reforms make this an energy-market issue, not just a facilities issue. [23]

Taken together, the findings support this concluding statement: a green data centre under peak load is not merely efficient—it is flexible, forecastable, thermally intelligent and grid-aware. Operators that pursue all four qualities will reduce energy use more credibly than those that chase a single annual efficiency number. [22]

7. Appendix and references

Abbreviations and acronyms

Term	Meaning
AI	Artificial intelligence
CUE	Carbon usage effectiveness
DER	Distributed energy resource
DR	Demand response
DRUPS	Diesel rotary uninterruptible power supply
ERE	Energy reuse effectiveness
HPC	High-performance computing
IT	Information technology load
KPI	Key performance indicator
PUE	Power usage effectiveness
RL	Reinforcement learning
SLA	Service-level agreement
UPS	Uninterruptible power supply
UTES	Underground thermal energy storage
WUE	Water usage effectiveness

Units used

Quantity	Unit
Instantaneous power	W, kW, MW
Electrical energy	kWh, MWh, TWh
Apparent electrical capacity	kVA
Temperature	°C

Refrigeration capacity ton, ton-hour

References

1. International Energy Agency, Energy and AI, Paris: IEA, 2025.
2. Google, "A new milestone for smart, affordable electricity growth: Data center demand response," March 19, 2026.
3. Uptime Institute, Uptime Institute Global Data Center Survey 2024, 2024.
4. International Energy Agency, "AI is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works," April 10, 2025.
5. Google DeepMind, Evans and Gao, "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%."
6. Google Cloud Blog, "Using demand response to reduce data center power consumption."
7. Google, "A new milestone for smart, affordable electricity growth; 1GW of data-centre demand response."
8. Radovanovic et al., "Carbon-Aware Computing for Datacenters," IEEE Transactions on Power Systems, 7 (3), 129–151, DOI: 10.1109/TPWRS.2022.3173250.
9. CenterSquare / Better Buildings, "Data Center Cooling Optimization Saves Energy and Cuts Costs in Mesa, AZ."
10. NTT Data, "AI in action: Saving energy, cutting costs and keeping data centers cool."
11. Microsoft Datacenters, "Measuring energy and water efficiency for Microsoft datacenters."
12. Meta Engineering, "Simulator-based reinforcement learning for data center cooling optimization."
13. Berkeley Lab / CII / IGBC / NetApp, "NetApp Bangalore Optimizes Data Center Cost Saving Through 'Free' Outside Air Cooling."
14. The Green Grid, "PUE: A Comprehensive Examination of the Metric."
15. Ana Radovanovic, Ross Koningstein, "Carbon-Aware Computing for Datacenters," Arxiv, June 2021.
16. Varun Mehra, Raiden Hasegawa, "Supporting power grids with demand response at Google data centers," October 2023.
17. CenterSquare / Better Buildings, "Data center cooling optimization saves energy and cuts costs in Mesa, AZ."
18. National Laboratory of the Rockies, "High-Performance Computing Data Center Power Usage Effectiveness."
19. Vertiv, "Summarizing the ASME paper: Power Usage Effectiveness Analysis of a High-Density Air-Liquid Hybrid Cooled Data Center."
20. Gheni, Kerskes and Stergiaropoulos, "Operational analysis of the cooling system in a direct liquid-cooled data center," Applied Energy, 2026.
21. Michael Terrell, "A new milestone for smart, affordable electricity growth," March 2026.
22. U.S. DOE / Lawrence Berkeley National Laboratory, DOE Data Center Load Flexibility Workshop Summary, 2025.
23. European Union, Delegated Regulation (EU) 2024/1364 on Data-Centre Reporting, 2024.