ISRG Journal of Engineering and Technology (ISRGJET)

OPEN ACCESS

ISRG PUBLISHERS

Abbreviated Key Title: ISRG J Eng Technol

ISSN: 3107-5894 (Online)

Journal homepage: https://isrgpublishers.com/isrgjet/

Volume – II Issue-II (March-April) 2026

Frequency: Bimonthly

AI-Powered Data Center Power Monitoring: Architectures, Algorithms, and Operational Intelligence

Ishak H.A MEDDAH

University of Saida Dr MOULAY Tahar, Saida, Algeria

Published: 18-04-2026

***Corresponding author: Ishak H.A MEDDAH**

Abstract

The exponential growth of digital services and artificial intelligence (AI) workloads has positioned data centers as critical infrastructure while simultaneously rendering them significant contributors to global energy consumption. Traditional monitoring and control paradigms, reliant on static thresholds and reactive rule-based systems, are increasingly inadequate for managing the dynamic complexity of modern facilities. This article presents a comprehensive examination of AI-driven methodologies for data center power monitoring and optimization. We synthesize contemporary research across multiple domains: machine learning-based power consumption modeling, non-intrusive workload disaggregation, predictive cooling control, and grid-interactive flexible operations. Drawing upon recent advances in deep learning, anomaly detection, and reinforcement learning, we analyze how AI transforms raw telemetry into actionable intelligence. The article evaluates architectural frameworks, algorithmic approaches including LSTM networks, Transformer models, and Isolation Forests, and presents validation results from production deployments. We demonstrate that integrated AI monitoring systems can reduce cooling energy by 15 40%, improve Power Usage Effectiveness (PUE) by up to 15%, and enable dynamic power shedding of 25 40% during grid stress events while preserving critical workload performance. Furthermore, we examine emerging metrics such as Power Compute Effectiveness (PCE) and the role of AI-driven digital twins in optimizing behind-the-meter renewable energy integration. Finally, we discuss implementation considerations, explainability requirements, and future research directions toward autonomous, self-optimizing data center infrastructures.

Keywords: *Data Center Monitoring, Artificial Intelligence, Machine Learning, Power Usage Effectiveness, Anomaly Detection, Predictive Control, Workload Flexibility, Power Compute Effectiveness, Digital Twins, Grid-Interactive Data Centers*

Copyright © ISRG Publishers. All rights Reserved.

DOI: 10.5281/zenodo.19638923

1. Introduction

1.1 The Data Center Energy Challenge

Data centers constitute the physical foundation of the digital economy, enabling cloud computing, streaming media, scientific research, and the burgeoning ecosystem of artificial intelligence applications. The International Energy Agency (IEA) estimates that global data center electricity consumption reached approximately 460 terawatt-hours (TWh) in 2024, representing 1.5-2% of worldwide electricity use, with projections suggesting potential doubling by 2030 driven primarily by AI workload proliferation (Power Engineering, 2025).

This consumption pattern presents multifaceted challenges. Operationally, energy costs represent a substantial portion of data center total cost of ownership. Environmentally, the carbon footprint of data center operations has attracted regulatory scrutiny and corporate sustainability commitments. Strategically, power availability has emerged as the primary constraint on data center expansion, with grid connection queues extending years in major markets and utilities struggling to accommodate both electrification trends and data center growth (Power Engineering, 2025). The IEA's newly launched Energy and AI Observatory aims to address the critical data gap in this area, providing policymakers with reliable information to navigate this fast-moving sector (Power Engineering, 2025).

1.2 Limitations of Conventional Monitoring

Traditional data center monitoring evolved from facility management and enterprise IT operations. Supervisory Control and Data Acquisition (SCADA) systems track environmental parameters, while Data Center Infrastructure Management (DCIM) platforms aggregate power distribution unit (PDU) and uninterruptible power supply (UPS) metrics. These systems share common limitations:

Reactive Thresholding: Conventional monitoring relies on static warning and alarm thresholds—for example, temperature exceeding 27°C or rack power approaching 80% of circuit breaker rating. These thresholds cannot adapt to changing operational patterns, seasonal variations, or equipment degradation. A Futurum survey shows that 67% of IT teams now use automation for monitoring, with zero respondents reporting no modern automation in their environment, signaling a decisive shift away from these static methods (Schmitt, 2026).

Siloed Data Architecture: Power monitoring, cooling system telemetry, and IT workload metrics typically reside in separate databases with limited integration, obscuring the causal relationships between computational activity and energy consumption. The ideal of Integrated Data Center Management (IDCM) remains unrealized in most facilities (Schmitt, 2026).

Temporal Coarseness: Many facilities collect data at 5-15 minute intervals, insufficient for capturing the burst characteristics of modern AI training workloads, which can create power demand swings of 30-50% within seconds.

Descriptive Rather Than Predictive: Traditional analytics answer "what happened?" but cannot reliably forecast future conditions or prescribe preventive actions.

1.3 The AI Opportunity

Artificial intelligence and machine learning offer transformative capabilities for data center power monitoring. Unlike rigid rule-based systems, AI models can:

- Learn complex, non-linear relationships between workload characteristics, environmental conditions, and power consumption
- Detect subtle anomalies preceding equipment failures or thermal excursions
- Forecast power demand with sufficient accuracy and lead time for proactive control
- Optimize control decisions across multiple objectives—energy efficiency, thermal safety, workload performance—in real-time
- Disaggregate aggregate measurements into component-level insights without intrusive instrumentation
- Enable natural language interfaces for infrastructure management, allowing operators to query system state conversationally (Introl, 2026)

This article systematically examines these capabilities, providing both theoretical foundations and practical implementation guidance for data center operators, sustainability officers, and infrastructure architects.

2. Foundational AI Techniques for Power Monitoring

2.1 Taxonomy of Monitoring Objectives

AI applications in data center power monitoring address distinct but interrelated objectives:

- Power Prediction: Forecast future power consumption at server, rack, or facility level. Key techniques: LSTM, GRU, Transformer, XGBoost. Output metrics: MAPE < 5%, RMSE.
- Anomaly Detection: Identify unusual patterns indicating inefficiency or failure risk. Key techniques: Isolation Forest, LOF, Autoencoders, ECOD. Output metrics: Precision, Recall, F1.
- Power Disaggregation: Estimate per-application power from aggregate measurements. Key techniques: NILM algorithms, Sparse Coding, Deep Learning. Output metrics: NMAE < 10%.
- PUE Optimization: Minimize facility overhead ratio through coordinated control. Key techniques: Reinforcement Learning, Bayesian Optimization. Output metrics: PUE reduction %.
- Grid Flexibility: Modulate consumption in response to grid signals. Key techniques: Workload scheduling, Power capping, RL. Output metrics: Demand reduction %.
- Power Quality Management: Mitigate harmonic distortion from GPU workloads. Key techniques: Classification algorithms, Real-time compensation. Output metrics: IEEE 519 compliance.

2.2 Core Algorithmic Approaches

2.2.1 Supervised Learning for Power Modeling

Power consumption modeling forms the foundation of AI monitoring. While early work employed linear regression relating

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

CPU utilization to server power, contemporary research demonstrates the necessity of multivariate approaches incorporating memory accesses, I/O activity, and thermal states. The shift to GPU-accelerated AI workloads has introduced new complexities, as these systems exhibit highly dynamic power profiles and introduce power quality challenges including harmonic distortion that must be managed according to standards like IEEE 519-2022 (Wu et al., 2025).

XGBoost and Gradient Boosting: Tree-based methods excel at capturing non-linear feature interactions with moderate data requirements. Racedo et al. (2025) demonstrated that XGBoost achieves mean absolute percentage error (MAPE) below 3% for server-level power prediction when trained on comprehensive performance counter data, though model generalization across hardware generations requires retraining.

Neural Network Architectures: Deep learning enables temporal modeling crucial for capturing workload dynamics. Convolutional neural networks (CNNs) extract local patterns from power time-series, while recurrent architectures (LSTM, GRU) maintain memory of historical context. Hybrid CNN-LSTM models have demonstrated particular efficacy for multi-step forecasting, capturing both short-term fluctuations and longer-term trends (Gebreyesus, 2025).

Attention Mechanisms and Transformers: The Transformer architecture, originally developed for natural language processing, has been adapted for time-series forecasting with promising results. Self-attention mechanisms enable the model to weigh the relevance of different historical timepoints, potentially capturing periodic patterns (daily, weekly cycles) without explicit feature engineering.

2.2.2Unsupervised Anomaly Detection

Data centers generate massive telemetry streams, making manual threshold configuration impractical. Unsupervised anomaly detection algorithms identify outliers without requiring labeled failure data. ServiceNow's event management demonstrates the power of this approach, reducing alert noise by 99% by clustering related alerts and surfacing only actionable insights (Introl, 2026).

Isolation Forest: This algorithm isolates anomalies through random feature partitioning, exploiting the property that anomalies are "few and different" requiring fewer partitions for isolation. Volkovičs (2025) applied Isolation Forest to Oracle Enterprise Manager metrics, demonstrating its effectiveness for detecting performance anomalies preceding power excursions.

Local Outlier Factor (LOF): LOF identifies anomalies based on local density deviation, detecting points whose neighbors are significantly denser than themselves. This local perspective proves valuable for identifying emerging hot spots or cooling failures before global thresholds trigger.

Empirical Cumulative Distribution Functions (ECOD): A more recent approach, ECOD computes outlier scores based on empirical distribution tails for each feature, assuming anomalies reside in distribution extremes. Its computational efficiency makes it suitable for real-time monitoring of high-dimensional sensor data.

2.2.3Reinforcement Learning for Control

Reinforcement learning (RL) provides a framework for learning optimal control policies through interaction with the environment. An RL agent observes the system state (temperatures, power, workload), takes actions (adjusting cooling setpoints, capping server power), and receives rewards based on outcomes (energy savings balanced against constraint violations).

Deep Q-Networks (DQN): DQN approximates the optimal action-value function using neural networks, enabling the agent to learn from high-dimensional state spaces. Chauhan (2025) applied DQN to HVAC control, achieving 15-25% cooling energy reduction while maintaining ASHRAE thermal guidelines.

Policy Gradient Methods: These methods directly optimize the control policy, often proving more stable for continuous action spaces like fan speed modulation or temperature setpoint adjustment.

Model-Based RL: Incorporating a learned dynamics model allows the agent to plan actions through simulated rollouts, potentially reducing real-world exploration costs. The PULSE framework (2025) integrates deep learning predictors with simulation to enable safe policy optimization.

2.2.4Large Language Models for Operations

The integration of large language models (LLMs) into AIOps platforms represents a significant advancement in human-AI collaboration. Research analyzing 183 articles published between January 2020 and December 2024 shows growing sophistication in applying language models to operational challenges (Introl, 2026).

Natural Language Interfaces: Modern AIOps platforms support chatbot- or LLM-powered interfaces allowing operators to query infrastructure state conversationally. The LLM translates natural language into appropriate monitoring queries and synthesizes results into comprehensible summaries. Smaller models like Mistral Small 7B demonstrate notable efficiency in reasoning and tool selection despite reduced size, making them suitable for edge deployment (Introl, 2026).

AI Agents for Autonomous Operations: ServiceNow's AI Agents for AIOps autonomously triage alerts, assess business and technical impact, investigate root causes, and drive remediation through coordinated agentic workflows. These agents represent a fundamental expansion from detection to autonomous remediation, handling routine incidents without human intervention (Introl, 2026).

3. AI Monitoring Architecture

3.1 Sensor Infrastructure and Data Acquisition

Effective AI monitoring begins with comprehensive data collection. The architecture must capture:

IT Layer Metrics:

- Server-level power consumption (via onboard sensors or PDU measurements)
- CPU utilization, memory bandwidth, I/O activity per server
- GPU utilization and memory for accelerated nodes, with attention to the unique power quality signatures of GPU workloads (Wu et al., 2025)
- Workload characteristics (job types, priorities, scheduling states)

Facility Layer Metrics:

- Rack inlet/outlet temperatures (multiple points per rack)
- CRAC/CRAH supply and return temperatures

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

- Chiller plant parameters (compressor status, chilled water temperatures)
- Pump and fan speeds, damper positions
- Weather data (outdoor temperature, humidity)
- Liquid cooling loop parameters: flow rates, temperatures, and leak detection for direct-to-chip or immersion cooling systems (Talib, 2026)

Grid Interface Metrics:

- Utility power import (kW, kVAR)
- On-site generation output (solar, fuel cells)
- Battery energy storage system (BESS) state of charge
- Grid signals (frequency, demand response requests)

Power Quality Metrics:

- Harmonic distortion measurements per IEEE 519-2022 standards
- Transient events and voltage sags
- Power factor variations

Data Quality Considerations: AI model performance depends critically on data quality. Key considerations include synchronization across heterogeneous time-series sources, handling of missing values (forward-filling, interpolation, or model-based imputation), detection and filtering of sensor faults, and appropriate aggregation intervals balancing resolution and data volume.

3.2 Edge-to-Cloud Processing Pipeline

Modern AI monitoring employs hierarchical processing:

- Edge Processing: Local aggregation nodes collect high-frequency data (1-second to 1-minute intervals) from power distribution units, server management controllers, and environmental sensors. Edge analytics perform real-time anomaly detection and trigger immediate alerts for critical conditions.
- Fog Layer: Regional aggregators (per data hall or facility) combine edge streams, perform temporal alignment, and execute forecasting models requiring broader context.
- Cloud/Central Platform: Historical data warehouses enable model training, what-if analysis, and cross-facility optimization. Centralized dashboards provide operator visibility and reporting for sustainability compliance.

3.3 Digital Twins for Simulation and Training

Digital twins—virtual representations of physical data center infrastructure—have emerged as essential tools for AI monitoring development and operational optimization (Schmitt, 2026). A comprehensive digital twin incorporates:

- Thermal dynamics models capturing airflow patterns, heat transfer, and equipment thermal inertia
- Power flow models representing distribution losses, UPS efficiency curves, and PDU loading
- Workload models simulating computational demands and scheduling behavior
- Control logic mirroring actual building management system (BMS) sequences
- Renewable energy integration models for behind-the-meter deployments (Data Center Dynamics, 2026a)

Digital twins enable safe training of reinforcement learning agents without risking physical equipment; what-if analysis for capacity planning and efficiency interventions; operator training in realistic scenarios; validation of AI recommendations before deployment; and testing of behind-the-meter configurations where renewable generation directly powers AI workloads (Data Center Dynamics, 2026a).

4. AI Applications in Power Monitoring

4.1 Predictive Power Consumption Modeling

Accurate power prediction enables proactive capacity management, efficient UPS sizing, and grid interaction optimization. Recent advances demonstrate significant capability improvements:

Server-Level Modeling: The comparative study by Racedo et al. (2025) evaluated multiple approaches on a testbed of heterogeneous physical machines. Their enhanced polynomial regression formula, incorporating CPU frequency, memory accesses, and disk I/O, achieved MAPE of 2.8% across diverse workloads, comparable to XGBoost (2.5%) while requiring substantially less training data and offering interpretable coefficients.

Facility-Level Forecasting: Gebreyesus (2025) developed hybrid CNN-LSTM models for 15-minute ahead prediction of total facility power, incorporating IT load, cooling parameters, and weather forecasts. The models achieved MAPE below 4% on production data from the ENEA CRESCO6 HPC facility, significantly outperforming ARIMA baselines.

Workload-Aware Prediction: The PULSE framework (2025) integrates deep learning predictors with workload scheduling information, enabling accurate forecasting under changing job mixes. This workload-awareness proves critical for AI-focused facilities where training jobs create highly variable demand patterns.

4.2 Anomaly Detection for Efficiency and Reliability

AI-powered anomaly detection identifies deviations from expected power behavior that may indicate equipment degradation, control faults, or efficiency opportunities.

Efficiency Anomalies: Models trained on normal PUE relationships can detect when cooling power increases disproportionately to IT load—potentially indicating refrigerant charge loss, fouled coils, or sensor drift. Volkovičs (2025) demonstrated that Isolation Forest and LOF algorithms detect such efficiency anomalies 2-4 weeks before they would trigger conventional thresholds, enabling proactive maintenance.

Thermal Anomalies: Local outlier detection identifies developing hot spots before they reach critical temperatures. By modeling spatial temperature distributions, AI systems distinguish between expected variations (higher loads in certain racks) and genuine anomalies (blocked perforated tiles, failing fans).

Power Quality Anomalies: High-frequency power monitoring

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

(sub-second sampling) enables detection of harmonics, sags, and transients that stress equipment and reduce efficiency. Machine learning classifiers trained on power quality signatures can identify problematic equipment or upstream grid disturbances, ensuring compliance with IEEE 519-2022 standards for harmonic distortion (Wu et al., 2025).

4.3 Non-Intrusive Power Disaggregation

Understanding which applications drive power consumption enables accurate carbon accounting, chargeback, and optimization prioritization. However, direct per-application power measurement requires OS-level instrumentation often unavailable in multi-tenant environments.

WattScope Architecture: Guan et al. (2024) developed WattScope, a system for non-intrusive application-level power disaggregation. The key insight exploits workload characteristics observed in production traces: low power variability, limited magnitude range, and high periodicity. These properties make disaggregation tractable where general-purpose blind source separation would fail.

WattScope adapts non-intrusive load monitoring (NILM) techniques from building energy research, applying them to server- and rack-level power measurements already available in data centers. The system employs factorial hidden Markov models and deep learning to estimate per-application power contributions without accessing guest operating systems. Evaluation on production Google Borg workloads demonstrates normalized mean absolute error (NMAE) consistently below 10%—sufficient for carbon accounting and trend analysis (Guan et al., 2024).

4.4 Predictive HVAC and Cooling Optimization

Cooling systems typically consume 30-40% of total data center energy, making them prime optimization targets. AI-driven predictive control consistently demonstrates 15-40% cooling energy reduction while maintaining or improving thermal compliance.

Google DeepMind Deployment: The pioneering DeepMind application at Google data centers achieved 40% cooling energy reduction, translating to a 15% decrease in overall PUE (Schmitt, 2026). The system used two years of historical monitoring data to train a neural network with 5 hidden layers and 50 nodes each, processing 19 normalized input variables. Every five minutes, the system pulls snapshots from thousands of sensors, feeds them through deep neural networks, and identifies actions minimizing energy consumption while satisfying safety constraints.

Schneider Electric and NVIDIA Partnership: In 2025, Schneider Electric partnered with NVIDIA to design AI-optimized reference architectures supporting rack densities up to 132 kW. The joint solution reduced cooling energy usage by nearly 20%, demonstrating vendor collaboration applying AI optimization to next-generation high-density infrastructure (Talib, 2026).

Reinforcement Learning Framework: Chauhan (2025) proposed an RL-based predictive control architecture integrating IoT sensor arrays providing real-time temperature, humidity, and pressure data; LSTM-based forecasting of near-term cooling demand; Deep Q-network agent learning optimal setpoint adjustments; and integration with building automation systems for autonomous execution. Simulation studies demonstrated 15-25% cooling energy savings relative to PID-based controls, with pilot deployments confirming practical feasibility.

4.5 Grid-Interactive Flexible Operation

As data center power demand grows, grid operators increasingly seek flexibility—the ability to modulate consumption during peak periods or renewable generation shortfalls. AI enables data centers to provide this flexibility without disrupting critical workloads.

Emerald AI Platform Trials: The Emerald Conductor platform has demonstrated AI-powered workload flexibility across multiple high-profile trials (Data Center Dynamics, 2026b; Adshead, 2026).

Phoenix, Arizona (May 2024): In collaboration with NVIDIA, Oracle Cloud Infrastructure, and utility Salt River Project, the system reduced AI cluster power consumption by 25% over three hours during a grid peak event while maintaining workload performance. Workloads were shifted between facilities with only 10ms latency penalties (Data Center Dynamics, 2026b).

Chicago Trial: Unlike the Phoenix demonstration where workload profiles were known beforehand, the Chicago test presented random, unknown workloads. The platform proved resilient in this significantly more challenging environment (Data Center Dynamics, 2026b).

UK National Grid Trial (December 2025): The most comprehensive test to date, conducted over five days at a Nebius facility near London with 96 Nvidia Blackwell Ultra GPUs, demonstrated:

- 40% demand reduction during extended events (up to 10 hours)
- 30% load shed within 30 seconds of simulated grid stress signals
- Successful reaction to spikes in demand during half-time at football matches
- Ability to help the grid navigate periods of low wind or extreme heat (Adshead, 2026)

Flexibility Mechanisms: AI systems orchestrate multiple flexibility mechanisms: workload slowing/pausing for flexible jobs; workload migration between facilities; intelligent workload tagging for priority-based orchestration; power capping (with performance tradeoffs); battery dispatch from on-site storage; and generator coordination for extended events.

Economic Value: Analysis suggests that if new AI data centers could flex consumption by 25% for two hours during less than 200 peak hours annually, this could unlock over 2 GW of additional grid connection capacity in the UK alone (Adshead, 2026). As Steve Smith, president of National Grid Partners, noted: "If you can throw more electrons at a fixed-cost system, you don't need to put more infrastructure in, and the rates come down for everyone else" (Adshead, 2026).

4.6 Behind-the-Meter Renewable Integration

The challenge of directly powering AI workloads with intermittent renewable energy has spurred innovative AI-driven solutions.

Soluna-Siemens Pilot: In January 2026, Soluna partnered with Siemens on a 2MW pilot project at the Project Grace site in Texas to address power management challenges for behind-the-meter data centers (Data Center Dynamics, 2026a). The project aims to validate an approach for managing rapid power demand fluctuations typical of GPU-driven AI workloads when directly powered by renewable energy.

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

The pilot integrates Siemens' electrical infrastructure, controls, and monitoring systems including transformers, switchgear, power converters, and Siemens SICAM SCADA platform for monitoring and control, with a structured commissioning process to document performance under variable compute demand conditions. The goal is to create a repeatable blueprint for future behind-the-meter deployments (Data Center Dynamics, 2026a).

4.7 Battery Energy Storage System (BESS) Optimization

AI-powered battery management has emerged as a critical capability for both backup reliability and grid interaction.

Narada Power Smart Cloud Platform: The platform developed by Narada Power demonstrates advanced AI integration for data center backup power (Narada Power, 2025). Key capabilities include:

- AI Inspection Engine: Real-time collection of cell voltage, temperature, and internal resistance data, with multi-physics coupling models enabling 24-hour advance risk detection and early warning
- Three-Dimensional Protection System: "cell-module-system" architecture providing multi-level fault isolation
- Precise SOC/SOH Estimation: Active balancing technology ensuring battery consistency and addressing capacity loss
- Multi-Dimensional Fault Diagnosis: Millisecond-level fault circuit interruption for conditions like micro-shorts and overcurrent, ensuring "zero propagation" of single-point failures

The platform also supports remote load circuit control, enabling zero standby energy consumption and dynamic allocation of backup resources based on load criticality—extending critical load backup time by over 200% while compressing non-critical load backup resources by 50% (Narada Power, 2025).

4.8 Explainable AI for Operator Trust

"Black box" AI recommendations face resistance from operators responsible for facility reliability. Explainable AI (XAI) techniques address this by providing transparency into model reasoning.

SHAP Values: SHapley Additive exPlanations (SHAP) assign credit for each prediction to input features based on cooperative game theory. Gebreyesus (2025) applied SHAP to energy prediction models, quantifying how specific features (IT load, outdoor temperature, fan speeds) influenced each forecast. This enables operators to understand why the AI recommends particular actions and builds confidence over time.

Visual Explanation Tools: Beyond numerical feature importance, effective XAI presents explanations visually—highlighting which time periods or sensor locations most influence predictions, or showing counterfactual scenarios ("if fan speed were 10% lower, predicted temperature would exceed threshold").

5. Emerging Metrics for the AI Era

5.1 Limitations of Traditional PUE

For nearly two decades, Power Usage Effectiveness (PUE) has been the primary metric for data center efficiency. However, AI-scale deployments are revealing structural considerations that PUE was not designed to address (Talib, 2026):

- PUE evaluates how efficiently a facility operates once energized, not how provisioned electrical capacity is allocated
- It does not assess how much capacity is structurally available for IT under declared redundancy configurations
- In regions with constrained utility interconnection, this structural availability is becoming more consequential than incremental overhead optimization

5.2 Power Compute Effectiveness (PCE)

Power Compute Effectiveness (PCE), developed by cooling system provider Airsys, addresses these limitations by providing transparency into power allocation within a data center's provisioned electrical envelope (Talib, 2026):

Definition: PCE is the ratio of provisioned IT compute power allocation to total provisioned site electrical capacity.

PCE does not measure real-time utilization or compete with PUE. Instead, it answers a structural question: Given a permitted power envelope and designed redundancy, how much of that envelope is sustainably allocatable to IT?

Practical Implications: In facilities where the provisioned electrical envelope is not fully utilized, PCE provides visibility into allocation. Operators may secure interconnection capacity and install infrastructure sized for future growth, yet only a portion actively supports IT load. PCE reveals this gap.

Cooling Architecture Trade-offs: As rack densities increase, liquid cooling technologies (direct-to-chip, rear-door heat exchangers, immersion) are moving from pilot to production scale (Talib, 2026). These approaches reduce fan energy but introduce pumping systems and heat rejection infrastructure that draw from the same provisioned capacity. Cooling architecture decisions directly influence PCE.

5.3 A Layered KPI Landscape

Dr. Rand Talib of Uptime Institute emphasizes that AI-scale infrastructure requires complementary metrics operating at different layers (Talib, 2026):

Metric	Question Answered	Layer
PUE	How efficiently does the facility operate once energized?	Operational efficiency
PCE	How much provisioned capacity is allocatable to IT?	Structural allocation
IT Utilization	How effectively is IT infrastructure utilized?	Compute productivity

A facility may exhibit strong PUE and PCE yet deliver poor compute productivity due to low IT infrastructure utilization. Conversely, a site may achieve high compute productivity but face structural expansion limits because cooling provisioning constrains allocatable IT capacity. Understanding AI infrastructure performance requires viewing these metrics together (Talib, 2026).

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

6. Implementation Framework

6.1 Maturity Model for AI Monitoring

Organizations implementing AI power monitoring typically progress through maturity stages:

Stage	Characteristics	Capabilities	Timeline
1: Foundational	Comprehensive metering, centralized data collection, basic dashboards	Descriptive analytics, manual analysis	6-12 months
2: Diagnostic	Historical data warehousing, automated reporting, threshold alerts	Root cause analysis, efficiency benchmarking	12-18 months
3: Predictive	Machine learning models, forecasting, anomaly detection	Proactive operations, predictive maintenance	18-24 months
4: Prescriptive	Optimization recommendations, what-if simulation	Decision support, efficiency opportunities	24-30 months
5: Autonomous	Closed-loop control, grid interaction, self-optimization	Dynamic flexibility, minimal human intervention	30-36 months

6.2 Technology Stack Selection

Data Platform Considerations:

- Time-series databases (InfluxDB, TimescaleDB, Prometheus) for metrics storage
- Stream processing (Apache Kafka, Apache Flink) for real-time analytics
- Data lake integration (cloud storage, Parquet format) for historical analysis

ML Operations (MLOps):

- Model versioning and experiment tracking (MLflow, DVC)
- Automated retraining pipelines triggered by data drift detection
- Model monitoring for prediction accuracy degradation
- A/B testing framework for comparing model versions

Integration Requirements:

- APIs for building management system (BMS) integration
- Read access to server management controllers (IPMI, Redfish)
- Workload scheduler integration (SLURM, Kubernetes, cloud APIs)
- Utility/grid operator communication interfaces (OpenADR, IEEE 2030.5)
- SCADA platform integration (e.g., Siemens SICAM) for behind-the-meter deployments (Data Center Dynamics, 2026a)

6.3 Model Lifecycle Management

Training Data Requirements: AI models require comprehensive training data covering normal operations and edge cases. For power prediction, at least 3-6 months of historical data is typically needed, including seasonal variations and diverse workload patterns. DeepMind's cooling optimization used two years of monitoring data (Schmitt, 2026). Anomaly detection may benefit from injected synthetic anomalies during training.

Validation and Testing: Rigorous validation protocols include:

- Temporal cross-validation respecting time-series dependencies
- Holdout periods representing unseen future conditions
- Stress testing with extreme scenarios (peak loads, equipment failures)
- Sensitivity analysis to input noise and missing data

Deployment Strategies:

- Shadow mode: Model runs in parallel with existing systems, recommendations logged but not acted upon
- Advisory mode: Recommendations presented to operators for approval
- Constrained autonomy: AI controls within bounded limits, human oversight for significant changes
- Full autonomy: Closed-loop optimization within safety constraints

6.4 Organizational Capabilities

Successful AI monitoring requires more than technology—it demands organizational readiness:

Cross-Functional Teams: Effective implementation brings together facility engineers (understanding physical constraints), IT operations (workload dynamics), data scientists (model development), and sustainability officers (goal definition).

Skill Development: Organizations must invest in training programs enabling operators to understand, trust, and effectively oversee AI systems. This includes fundamentals of machine learning concepts, interpretation of model outputs, and escalation procedures for anomalous situations.

Governance Processes: Clear policies should govern model updates, performance review cycles, and intervention protocols. Regular model audits ensure continued alignment with operational objectives and safety requirements.

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

7. Case Studies

7.1 ENEA CRESCO6 HPC Facility

The Italian National Agency for New Technologies, Energy, and Sustainable Economic Development (ENEA) operates the CRESCO6 high-performance computing facility. Gebreyesus (2025) implemented an AI-driven predictive framework addressing both energy and cooling efficiency.

Challenge: The facility needed to optimize cooling energy while maintaining thermal safety for dense HPC equipment, with complex interactions between IT load, cooling configuration, and external weather.

Solution: The research team developed a comprehensive framework including SHAP-assisted feature selection identifying key predictors; hybrid CNN-LSTM models for 15-minute temperature and energy forecasting; sensitivity analysis evaluating fan speed adjustments (±10% to ±50%); and XAI integration using SHAP values for model transparency.

Results: The CNN-LSTM models achieved superior accuracy for all prediction targets. Sensitivity analysis revealed that ±50% fan speed adjustments significantly impacted thermal dynamics, identifying zones requiring enhanced monitoring. The framework provided actionable insights for optimizing cooling operations while maintaining ASHRAE thermal guidelines.

7.2 Emerald AI Grid Flexibility Demonstrations

Phoenix, Arizona (May 2024): The Emerald Conductor platform orchestrated AI workloads across 256 NVIDIA GPUs, automatically modulating power consumption based on real-time grid conditions while preserving service quality. Results included 25% power reduction over three hours, maintained critical workload performance, and gradual consumption ramping (15-minute ramp) to avoid grid shocks (Data Center Dynamics, 2026b).

Chicago: Unlike Phoenix where workload profiles were known beforehand, the Conductor handled unknown, random workloads, proving resilient in a significantly more challenging environment (Data Center Dynamics, 2026b).

UK National Grid Trial (December 2025): The most comprehensive demonstration to date, conducted over five days at a Nebius facility near London with 96 Nvidia Blackwell Ultra GPUs, tested more than 200 real-time simulated grid events (Adshead, 2026). Key results included up to 40% demand reduction during extended events (up to 10 hours), 30% load shed within 30 seconds of simulated system stress, successful reaction to demand spikes during halftime at football matches, and demonstrated ability to help the grid navigate periods of low wind or extreme heat.

Varun Sivaram, CEO of Emerald AI, emphasized three mechanisms for achieving flexibility: slowing or pausing flexible workloads, moving workloads between datacenters with minimal latency impact, and intelligent workload tagging for priority-based orchestration (Adshead, 2026).

7.3 Presight-Khazna UAE Deployment

In October 2025, Presight signed an MOU with Khazna Data Centers to deploy an AI-optimized facility management system across Khazna's entire network of 30 data centres in the UAE (Teletimes International, 2025).

Scope: The project will implement an AI-powered command and control platform operating from a secure hub in Abu Dhabi, using artificial intelligence to monitor energy, cooling, equipment performance, and security across the network.

Capabilities: Predicting issues before they occur, optimizing operations 24/7, early fault detection to reduce downtime, smarter cooling systems to lower energy usage, predictive maintenance to limit unnecessary servicing, and unified oversight ensuring operational continuity.

Strategic Significance: The command and control centre will form part of G42's Intelligence Grid, a globally distributed network of AI infrastructure hubs, unifying management of sites from the UAE to Singapore, Kazakhstan, Europe, and Africa. As Thomas Pramotedham, CEO of Presight, stated: "With Khazna, we are bringing Applied Intelligence to the front lines of digital infrastructure, embedding AI at the core of mission-critical systems to deliver efficiency, resilience, and sustainability at scale" (Teletimes International, 2025).

7.4 Narada Power Smart Cloud Platform

Narada Power's development of an AI-driven Smart Cloud Platform demonstrates advanced battery energy storage system (BESS) optimization for data center applications (Narada Power, 2025).

Key Innovations: Three-dimensional protection ("cell-module-system" architecture), AI inspection engine for 24-hour advance risk detection, multi-dimensional fault diagnosis with millisecond-level interruption, and remote branch control for cloud-based load circuit management.

Results: Critical load backup time increased by over 200% through differentiated backup strategies; non-critical load backup resources compressed by 50%; active balancing technology ensures battery consistency and addresses capacity loss. The platform transforms traditional passive maintenance into active defense, moving toward "zero manual intervention" and redefining data center energy management reliability (Narada Power, 2025).

7.5 Soluna-Siemens Behind-the-Meter Pilot

Announced in January 2026, this 2MW pilot at Soluna's Project Grace site in Texas addresses the challenge of powering GPU-driven AI workloads directly with renewable energy (Data Center Dynamics, 2026a).

Objectives: Validate behind-the-meter approach for managing rapid power demand fluctuations, integrate Siemens electrical infrastructure, controls, and monitoring systems, document performance under variable compute demand conditions, and create a repeatable blueprint for future behind-the-meter deployments.

Equipment: Siemens will provide transformers, switchgear, power converters, and SICAM SCADA platform for monitoring and control. The project addresses one of the biggest challenges to scaling AI technology: the amount of compute power it demands and the grid constraints limiting new deployments (Data Center Dynamics, 2026a).

8. Challenges and Limitations

8.1 Technical Challenges

Data Quality and Availability: AI models require high-quality training data, yet many facilities lack comprehensive instrumentation. Legacy equipment may lack monitoring interfaces, and sensor drift or failures can corrupt model inputs.

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

Data synchronization across heterogeneous systems remains nontrivial. The IEA's Energy and AI Observatory was specifically created to address the limited data on AI data center energy consumption (Power Engineering, 2025).

Model Generalization: Models trained on specific hardware configurations may not transfer to new equipment generations. Server power characteristics vary significantly between manufacturers, and cooling dynamics differ across facility designs. Retraining requirements create operational overhead.

Temporal Dynamics: Data center operations evolve—workload patterns shift, equipment ages, and control strategies change. Models must adapt to concept drift through continuous learning mechanisms, raising questions about stability and safe exploration.

Safety Guarantees: Reinforcement learning agents exploring control actions could theoretically discover unsafe states. Ensuring safety requires careful reward design, action constraints, and fallback mechanisms that may limit optimization potential.

Power Quality Management: GPU-driven AI workloads introduce harmonic distortion and power quality challenges requiring active management to maintain compliance with IEEE 519-2022 standards (Wu et al., 2025).

8.2 Organizational Challenges

Skill Gaps: The intersection of facility engineering, IT operations, and data science is sparsely populated. Organizations struggle to find talent combining these domains, and existing staff may resist AI-driven changes to familiar workflows.

Trust and Transparency: Operators responsible for multi-million-dollar facilities rightfully hesitate to delegate control to black-box algorithms. Building trust requires explainability tools, gradual deployment, and demonstrated reliability over time. The integration of LLM-based natural language interfaces can help bridge this gap by making AI reasoning more accessible (Introl, 2026).

Cross-Functional Silos: Facility teams report through real estate or operations, while IT teams report through technology organizations. Budgets, incentives, and metrics often misalign, hindering integrated AI monitoring initiatives.

8.3 Economic Considerations

Investment Requirements: Comprehensive AI monitoring requires investment in sensors, data infrastructure, software platforms, and skilled personnel. Payback periods vary based on facility size, energy costs, and efficiency opportunities. The 40% cooling energy reduction achieved by DeepMind demonstrates the scale of potential return (Schmitt, 2026).

Value Quantification: Benefits extend beyond direct energy savings to include avoided downtime, extended equipment life, accelerated capacity deployment, and grid service revenues. Quantifying these diverse value streams challenges traditional ROI analysis.

Vendor Lock-in: Organizations adopting proprietary AI monitoring platforms may face switching costs and limited flexibility. Open standards and modular architectures mitigate this risk.

9. Future Directions

9.1 Foundation Models for Data Center Operations

The emergence of large language models and foundation models suggests potential for pre-trained "data center foundation models" capturing general knowledge of facility dynamics, equipment characteristics, and operational patterns. Such models could be fine-tuned for specific facilities with limited data, dramatically accelerating AI deployment. Research on LLM applications in AIOps, analyzing 183 articles published between 2020 and 2024, demonstrates growing sophistication in applying language models to operational challenges (Introl, 2026).

9.2 Multi-Facility Federated Learning

Data center operators managing global footprints could employ federated learning to train models across facilities without centralizing sensitive operational data. This enables learning from diverse conditions while respecting data sovereignty and security requirements. The Presight-Khazna deployment, which will unify management across the UAE and internationally through G42's Intelligence Grid, points toward this distributed intelligence model (Teletimes International, 2025).

9.3 Carbon-Aware Optimization

Beyond energy efficiency, future AI monitoring will optimize for carbon intensity, scheduling flexible workloads to times and locations with cleaner grid power. This requires integration with grid carbon forecasting and may involve coordinating across geographically distributed facilities. The Soluna-Siemens behind-the-meter pilot directly addresses this challenge for renewable-powered sites (Data Center Dynamics, 2026a).

9.4 Self-Healing Infrastructure

Advanced AI monitoring will not only detect anomalies but initiate automated remediation—rerouting workloads around failing hardware, adjusting cooling to compensate for failed fans, or gracefully degrading non-critical services during supply constraints. ServiceNow's AI Agents for AIOps already demonstrate autonomous triage and remediation capabilities, handling routine incidents without human intervention (Introl, 2026).

9.5 Hardware-AI Co-Design

Future servers and facility equipment will embed AI capabilities at design time, with onboard processors running monitoring models and communicating directly with facility-level optimization systems. This tight integration will enable faster response and finer granularity than current add-on approaches. The Schneider Electric-NVIDIA partnership on AI-optimized reference architectures represents an early step in this direction (Talib, 2026).

9.6 Grid-Interactive Data Centers as Standard Practice

Based on successful trials in Phoenix, Chicago, and the UK, grid-interactive flexible operation is moving from demonstration to commercial deployment. Emerald AI's first commercial project will deploy at NVIDIA's 96MW Aurora data center in Manassas, Virginia (Data Center Dynamics, 2026b). As National Grid's Steve Smith noted, the key is managing peak demand during the small number of hours when the grid is stretched, enabling more efficient use of existing infrastructure (Adshead, 2026).

10. Conclusion

The convergence of artificial intelligence and data center power monitoring represents a fundamental shift from reactive, threshold-based management to proactive, predictive, and ultimately autonomous operation. This article has synthesized contemporary research demonstrating that AI-powered monitoring delivers measurable benefits across multiple dimensions:

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923

- Accuracy: Machine learning models predict power consumption with MAPE below 5%, enabling confident capacity planning and grid interaction
- Efficiency: Predictive cooling control reduces HVAC energy by 15-40%, directly improving PUE, with DeepMind achieving 40% reduction (Schmitt, 2026)
- Reliability: Anomaly detection identifies developing issues weeks before conventional alarms, with platforms like ServiceNow reducing alert noise by 99% (Introl, 2026)
- Flexibility: AI workload orchestration enables 25-40% demand reduction during grid stress, accelerating interconnection and supporting renewable integration, as demonstrated in multiple Emerald AI trials (Adshead, 2026; Data Center Dynamics, 2026b)
- Structural Visibility: Emerging metrics like PCE provide transparency into capacity allocation within fixed power envelopes, addressing constraints that PUE cannot capture (Talib, 2026)
- Behind-the-Meter Integration: AI enables management of rapid power fluctuations when data centers are directly powered by renewable energy, reducing grid dependence (Data Center Dynamics, 2026a)

The path to widespread adoption requires continued advances in model robustness, safety guarantees, and integration standards. Organizations must invest not only in technology but in cross-functional teams, skill development, and governance processes enabling effective human-AI collaboration. Major deployments like the Presight-Khazna network-wide AI platform in the UAE demonstrate that this transition is already underway at scale (Teletimes International, 2025).

As data centers continue their evolution from passive computing facilities to active grid participants and self-optimizing infrastructures, AI monitoring will transition from competitive advantage to operational necessity. The facilities that succeed in this transition will deliver not only lower costs and reduced environmental impact but also the reliability and scalability that the digital economy demands. With the IEA projecting data center electricity consumption could more than double by 2030, the role of AI in managing this growth sustainably has never been more critical (Power Engineering, 2025).

References

1. Adshead, A. (2026). National Grid, Nebius and Emerald hail datacentre power throttling. *Computer Weekly*, March 2026.
2. Chauhan, M. (2025). AI-Driven Predictive Control for Data Center HVAC Systems. *Heat Pumping Technologies Magazine*, 43(3).
3. Data Center Dynamics. (2026a). Soluna partners with Siemens on 2MW pilot to test power management challenges for behind-the-meter data centers. *Data Center Dynamics*, January 2026.
4. Data Center Dynamics. (2026b). Emerald AI releases results of UK demonstration project in partnership with National Grid. *Data Center Dynamics*, March 2026.
5. Gebreyesus, Y. (2025). AI-Driven Techniques for Energy and Cooling Efficient Data Centre Management. *University College Dublin Doctoral Thesis*.
6. Guan, X., Bashir, N., Irwin, D., & Shenoy, P. (2024). WattScope: Non-intrusive Application-level Power Disaggregation in Datacenters. *ACM SIGMETRICS Performance Evaluation Review*, 51(4), 24-25.
7. Introl. (2026). AIOps for Data Centers: Using LLMs to Manage AI Infrastructure. *Introl Blog*, January 2026.
8. Narada Power. (2025). AI赋能“数字神经” 南都电源打造数据中心Smart云平台. *Narada Power News*, October 2025.
9. Power Engineering. (2025). IEA launches observatory to monitor AI and data center energy demand. *Power Engineering*, June 2025.
10. PULSE: A modular framework for predictive energy efficiency in heterogeneous data centers. (2025). *SoftwareX*, 31, 102313.
11. Racedo, S., Jaumard, B., Glatard, T., Delgado, O., & Masoudi, M. (2025). Advances in power consumption model for data centers: Analytical formulas vs. machine learning models. *Future Generation Computer Systems*.
12. Schmitt, M. (2026). Data Center Monitoring, Management, and Control: Stacking it All Up for the AI Era. *Nlyte Blog*, January 2026.
13. Talib, R. (2026). Capacity allocation and the next generation of AI-era KPIs. *Uptime Intelligence Update 477*, March 2026.
14. Teletimes International. (2025). Presight to deploy AI-driven facility management system across Khazna data centre network. *Teletimes International*, October 2025.
15. Volkovičs, R. (2025). Data Centre Monitoring Model Utilizing Artificial Intelligence, Machine Learning and Anomaly Detection Algorithms. *16th International Scientific and Practical Conference “Environment. Technology. Resources”*.
16. Wu, R., et al. (2025). AI-Driven Data Center Energy Profile, Power Quality, Sustainable Sitting, and Energy Management: A Comprehensive Survey. *2025 IEEE Conference on Technologies for Sustainability (SusTech)*, pp. 1-8.

Copyright © ISRG Publishers. All rights Reserved.
DOI: 10.5281/zenodo.19638923