Samuel N Nimaful, 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Energy Efficiency Optimization in Hyperscale Data Centers

Samuel N Nimaful, Joel Holison, Gloria O. Darkoh, Augustine Hanyabui, Faith Esther Holison, Laureta Tatenda Nyamsutswa
Eastern Illinois University

Abstract- Hyperscale data centers—very large, industrialized facilities that deliver cloud-scale compute, storage, and networking—have become a central node in the global energy system because they concentrate electrical load, water demand, and waste heat while enabling rapidly growing digital services and AI workloads. Over the past decade, efficiency gains in hyperscale infrastructure and IT operations have helped decouple growth in compute demand from the growth in electricity use, but the acceleration of GPU/accelerator-based AI is now reintroducing strong upward pressure on both energy and cooling capacity requirements (Masanet et al., 2020; International Energy Agency [IEA], 2025; U.S. Department of Energy [DOE], 2024).[1]

Keywords: Hyperscale Data Centers, Cloud Computing Infrastructure, Artificial Intelligence (AI) Workloads, GPU/Accelerator-Based Computing, Energy Consumption.

I. INTRODUCTION

Recent global assessments indicate that data centres consumed about 415 TWh in 2024 (≈1.5% of global electricity), and under a base-case projection could rise toward ~945 TWh by 2030 driven largely by accelerated computing for AI (IEA, 2025).[2] In the United States, a DOE-sponsored Lawrence Berkeley National Laboratory update estimates U.S. data center electricity consumption rose from 58 TWh (2014) to 176 TWh (2023) and could reach 325–580 TWh by 2028 depending on scenarios; DOE's summary translates this to ~6.7%–12% of total U.S. electricity by 2028 (Shehabi et al., 2024; DOE, 2024).[3] These projections highlight why hyperscale efficiency must be treated as a multi-layer optimization problem spanning (a) facility thermodynamics and heat rejection, (b) electrical conversion and distribution losses, (c)

IT utilization and "energy proportionality," (d) workload placement and scheduling (including carbon-aware scheduling), and (e) governance and reporting through standardized metrics (e.g., PUE, WUE, CUE, ERF) and emerging regulatory schemes (International Organization for Standardization [ISO], 2016–2023; European Commission, 2024).[4]
Across the literature (2016–2026) and leading operator disclosures, several conclusions consistently emerge. First, PUE improvements in hyperscale have approached practical lower bounds in many mature sites (often around ~1.1–1.2), meaning further large gains increasingly require either (i) moving beyond air cooling toward liquid/immersion and higher-temperature heat reuse, or (ii) shifting the definition of "efficiency" toward metrics of useful work per energy and carbon/water intensity (Masanet et al., 2020; ISO, 2016–2023; Uptime Institute, 2023–2024).[5] Second, the "AI era" is pushing rack and chip power densities upward, which makes liquid cooling (direct-to-chip, rear-door heat exchangers, or immersion) strategically important because it enables higher heat capture, reduced fan energy, extended economizer hours, and higher-quality waste heat for reuse (ASHRAE TC 9.9, 2021; ISO, 2023; Schneider Electric, 2025; Shehabi et al., 2024).[6]

Third, facility-side efficiency cannot compensate for persistently low server utilization; therefore, virtualization, consolidation, and workload-aware scheduling remain among the highest-leverage measures because they reduce IT energy directly and also reduce "secondary" cooling and power provisioning needs (DOE, 2024; Masanet et al., 2020).[7] Fourth, carbon management is shifting from annual renewable matching toward hourly (24/7) carbon-free energy matching, which makes spatio-temporal load shifting, demand response, and grid-interactive data centers an emerging

frontier that couples IT scheduling with power system constraints (IEA, 2025; Wiesner et al., 2021; Riepin et al., 2025; Google, 2023).[8]

This paper synthesizes peer-reviewed research, government and standards documents, and primary operator/vendor disclosures to provide a publishable, metrics-grounded framework for hyperscale efficiency optimization. It contributes: (1) a practical taxonomy linking metrics (PUE/CUE/WUE/ITEE/ITEU/CER/ERF) to engineering levers; (2) comparative tables of cooling, power, and IT optimization options; (3) operator case-study outcomes using publicly reported metrics; and (4) quantification templates for energy, water, carbon, life-cycle assessment (LCA), and cost-benefit analysis aligned with ISO and GHG Protocol guidance (ISO, 2006; Greenhouse Gas Protocol, 2015).[9]

II. METHODOLOGY

This research uses a structured narrative review approach designed for publishable synthesis rather than a full PRISMA-style systematic review, because comprehensive database export and deduplication workflows are not available in this environment. Nevertheless, the search and screening process was designed to be transparent, reproducible, and biased toward primary sources.

Search window and timeframe. Searches were executed and sources screened in March 2026 (U.S. Eastern time). The primary review window targeted 2016–2026 to satisfy the "last 10 years" requirement, while selectively including older foundational works (e.g., energy-proportional computing) when necessary to explain core concepts that remain causal drivers of modern strategies (Barroso & Hölzle, 2007; Masanet et al., 2020).[10]

Source types and prioritization. Sources were prioritized in this order: 1. Peer-reviewed research in high-quality venues (e.g., Science, Nature, Applied Energy, Renewable & Sustainable Energy Reviews, IEEE/ACM proceedings) addressing data center energy, cooling, workload optimization, carbon-aware scheduling, or LCA (Masanet et al., 2020; Lazic et al., 2018; Wiesner et al., 2021; Souza et al., 2023; Alissa et al., 2025).[11]

2. Government and national-lab reports and technical guides (e.g., DOE/LBNL, IEA, IEA 4E) providing bottom-up energy modelling, scenario ranges, and policy-relevant metrics (Shehabi et al., 2016; Shehabi et al., 2024; DOE, 2024; IEA 4E, 2025).[12]

3. Standards and technical guidance from recognized bodies (e.g., ISO/IEC 30134 KPI series; ASHRAE TC 9.9 guidelines and white papers; ASHRAE energy standard developments) to ensure definitions and boundaries are consistent (ISO, 2016–2023; ASHRAE TC 9.9, 2021).[13]

4. Primary operator disclosures and white papers from hyperscale cloud providers and major operators (e.g., PUE/WUE reporting pages, sustainability reports, methodology documents) for case studies and real-world metrics (Google, 2024; Amazon Web Services, 2024; Microsoft, 2024–2025; Meta, 2024).[14]

5. Vendor technical white papers when they provide engineering detail not otherwise available (e.g., direct liquid cooling integration, immersion requirements/specifications), with claims cross-checked against standards or academic consensus where possible (Open Compute Project, 2022; Schneider Electric, 2025).[15]

Search strategy (queries and chaining). Search queries combined four themes: (a) "hyperscale data center" definitions and fleet metrics; (b) data center KPIs (PUE, WUE, CUE, ERF, ITEEsv, ITEUsv, CER); (c) engineering technologies (liquid/immersion/free cooling, UPS and power distribution, renewable integration, heat reuse); and (d) algorithmic strategies (virtualization, consolidation, workload placement, carbon-aware scheduling, reinforcement learning for cooling). Representative query terms included: ISO/IEC 30134-2 PUE categories, ASHRAE TC 9.9 liquid cooling, direct-to-chip cooling energy, carbon-aware scheduling data center, data center energy usage report 2024 LBNL, and EU data centre reporting regulation 2024/1364. Backward and

forward citation-chaining was simulated by following references embedded in primary reports (e.g., Masanet et al., 2020 referencing LBNL series work; Shehabi et al., 2024 referencing thermodynamics-based PUE/WUE modelling).[16]

Inclusion and exclusion criteria. - Included: sources with (i) explicit methods, definitions, or measured outcomes; (ii) relevance to hyperscale or large facilities, or transferable engineering/IT principles; (iii) credible provenance (peer review, standards bodies, government labs, primary operator reporting). - Excluded or de-emphasized: purely marketing content without methodological detail; unverifiable claims; duplicated summaries of primary sources; or content focused solely on small enterprise facilities unless the principle generalizes.
Limitations. Operator metrics (especially water and carbon) remain non-uniform across the industry due to different boundary choices (site vs source water; location-based vs market-based electricity emissions; treatment of embodied emissions), which complicates cross-provider comparisons even when KPI names match. This is itself a key finding and is addressed explicitly in the metrics and carbon-accounting sections (GHG Protocol, 2015; Shehabi et al., 2024; European Commission, 2024).[17]

III. BACKGROUND AND DEFINITIONS

Hyperscale data centers as an energy system actor
While there is no single universal definition, "hyperscale" generally refers to data centers designed for extreme scalability, high automation, and large homogeneous deployments supporting cloud platforms and internet-scale services. Over time, the term has converged around facilities operated by a relatively small set of global firms, often built in modular expansions ("repeatable" design blocks) and optimized for high utilization and cost efficiency (Katal et al., 2022; Masanet et al., 2020).[18]

From an energy perspective, hyperscale matters because of (a) load size and clustering (multi-10s to 100s of MW per campus), (b) high capacity factors (near-continuous demand), (c) rapid load growth driven by AI accelerators, and (d) increasing interaction with power grids through renewable procurement, on-site generation/storage, and demand response (DOE, 2024; IEA, 2025).[19]

Core efficiency metrics and why definitions matter
Efficiency optimization in hyperscale data centers is highly sensitive to measurement boundaries. Metrics can be gamed unintentionally (or intentionally) if boundaries exclude key losses or externalities; therefore, modern best practice is to use standardized KPIs and disclose measurement categories.

Table 1 Summarizes The Most Relevant Kpi Family For Hyperscale Optimization, Grounded In The Iso/Iec 30134 Series And Related Industry Frameworks.

Kpi	What It Measures	Canonical Definition (Conceptual)	Typical Unit / Dimension	Why It Matters In Hyperscale
Pue (Power Usage Effectiveness)	Electrical Overhead Intensity	Total Facility Energy ÷ It Equipment Energy	Dimensionless	Captures Cooling + Power Delivery Overhead; Near Lower Bound In Leading Hyperscale, So Incremental Gains Are Harder And More Expensive
Wue (Water Usage Effectiveness)	Water Intensity Of It Energy	Water Used For Cooling/Humidification ÷ It Equipment Energy	L/Kwh	Makes Water Tradeoffs Explicit; Critical As Evaporative Cooling And Water Stress Become Constraints
Cue (Carbon Usage Effectiveness)	Operational Co_2 Intensity Of It Energy	Co_2 Emissions Associated With Operations ÷ It Equipment Energy	$Kgco_2E/Kwh$-It	Links Power Sourcing And Facility Efficiency; Supports Decarbonization Comparisons

Kpi	What It Measures	Canonical Definition (Conceptual)	Typical Unit / Dimension	Why It Matters In Hyperscale
Erf (Energy Reuse Factor)	Fraction Of Energy Reused Outside Boundary	Reused Energy ÷ Total Data Center Energy	Dimensionless	Incentivizes Waste-Heat Reuse (District Heating, Industrial Processes) When Feasible
Cer (Cooling Efficiency Ratio)	Cooling System Efficiency Kpi	Standardized Ratio For Energy Used To Control Temperatures In Dc Spaces	Kpi (Iso-Defined)	Supports Cooling Design Comparison Beyond Aggregate Pue
Iteesv (It Equipment Energy Efficiency – Servers)	"Useful Work" Per Server Energy	Benchmark-Based Efficiency Kpi For Servers	Dependent On Benchmark	Encourages Efficient Server Selection And Configuration
Iteusv (It Equipment Utilization – Servers)	Server Utilization Kpi	Aggregate Server Cpu Utilization Kpi	% (Conceptually)	Highlights Underutilization; Informs Consolidation, Scheduling, And Capacity Planning

Sources for Table 1: ISO/IEC 30134-2 (PUE categories), ISO/IEC 30134-8 (CUE), ISO/IEC 30134-9 (WUE), ISO/IEC 30134-6 (ERF), ISO/IEC 30134-7 (CER), ISO/IEC 30134-4 (ITEEsv), ISO/IEC 30134-5 (ITEUsv), and The Green Grid's foundational metric framing for carbon and water effectiveness.[20]

PUE categories and "near-floor" measurement

Even within PUE, measurement category changes what is counted as "IT energy." ISO/IEC 30134-2 defines multiple categories based on where IT energy is measured (e.g., UPS output vs power distribution unit output vs IT equipment input), which can materially alter comparability across sites (ISO, 2016).[21]

A practical implication is that hyperscale operators aiming for publishable transparency should report: (1) the PUE number, (2) the measurement boundary/category, (3) the averaging window (annual vs trailing twelve months), and (4) whether sites are in "stable operations" (Google, 2024; ISO, 2016).[22]

IT-side efficiency as a first-order driver

Facility efficiency cannot compensate for wasted IT energy. The field's foundational concept of energy-proportional computing—the idea that power should scale with utilization—remains relevant because servers often consume substantial idle power, creating a structural incentive for consolidation, virtualization, and turning unused capacity off (Barroso & Hölzle, 2007; Masanet et al., 2020).[10]

IV. LITERATURE REVIEW

Macro-trends in data center electricity use and the hyperscale efficiency effect

A landmark synthesis of global data center energy use found that from 2010 to 2018, global electricity use rose only modestly (≈6%) even as compute instances grew sharply, attributing decoupling to improved server efficiency, reduced idle power, higher virtualization, and lower PUE—trends driven substantially by migration of workloads into cloud/hyperscale facilities (Masanet et al., 2020).[23] However, newer assessments emphasize that the post-2017 era is characterized by increasing accelerator deployment, higher rack densities, and heightened regional grid impacts. The IEA's 2025 analysis states that data centres consumed ~415 TWh in 2024 and projects strong growth toward

2030, with the scale and location of incremental load creating policy urgency (IEA, 2025).[2]

In the U.S., a DOE-sponsored update estimates accelerating growth: stable energy use around 2014–2016 followed by rising consumption and a broad scenario range to 2028, reflecting uncertainty in AI adoption and infrastructure choices (Shehabi et al., 2024; DOE, 2024).[3]

Why estimates diverge and why it matters for optimization research

A critical review of data-center energy models reports wide divergence in published estimates and projections due to inconsistent boundaries (what counts as "data center" vs networks), mixed top-down vs bottom-up methods, and limited transparency from operators. Using company-level data and modelling, the review estimates global data centers used on the order of ~300–380 TWh in 2023 and highlights that a small number of large operators account for a substantial share of electricity use (IEA 4E, 2025).[24]

For research on hyperscale optimization, this divergence implies two needs: (1) publishable studies should specify boundaries and metric definitions (ISO/IEC 30134 alignment), and (2) optimization results should be expressed not only as PUE deltas but also as impacts on absolute electricity, water, and carbon under realistic utilization and growth scenarios (ISO, 2016; Shehabi et al., 2024).[25]

Cooling research themes: from economizers to liquid and immersion

Cooling has long represented a major share of "overhead" energy in traditional facilities, and modern hyperscale design increasingly focuses on reducing compressor-based refrigeration via economization, higher allowable temperatures, and airflow management (DOE, 2024; ASHRAE TC 9.9, 2021).[26]

Within the last decade, the literature has shifted from air-side and water-side economizers toward liquid cooling as a structural enabler for high-density compute. ASHRAE's TC 9.9 white paper documents the mainstreaming of liquid cooling, driven by rising chip power and densification (ASHRAE TC 9.9, 2021).[27] Peer-reviewed research emphasizes that direct liquid cooling enables higher cooling temperatures, which can reduce cooling energy and enable meaningful waste heat reuse potential (Stahlhut et al., 2025).[28]

Immersion cooling is supported by an expanding ecosystem of open specifications addressing fluid properties, compatibility, and system interfaces—important because supply chain standardization is a barrier to adoption at hyperscale scale (Open Compute Project, 2022).[29]

AI/ML in data center operations: control and scheduling

Two distinct AI/ML themes dominate the last decade:

Facility control optimization (cooling and airflow). Early deployments reported substantial reductions in cooling energy using ML-based control, with later peer-reviewed work demonstrating reinforcement-learning/model-predictive control applied to large-scale data center cooling regulation (DeepMind, 2016; Lazic et al., 2018).[30]

IT scheduling and carbon-aware computing. Research has matured from energy-aware consolidation to explicit carbon optimization that incorporates variable grid carbon intensity and latency constraints across geo-distributed data centers (Wiesner et al., 2021; Souza et al., 2023; Riepin et al., 2025).[31]

The implication for hyperscale efficiency is that "optimization" increasingly means joint optimization across layers, where facility telemetry and grid signals inform workload placement, and workloads are shaped to participate in demand response or 24/7 carbon-free matching (Google, 2023; IEA, 2025).[32]

Governance and regulation: transparency is becoming mandatory in some regions

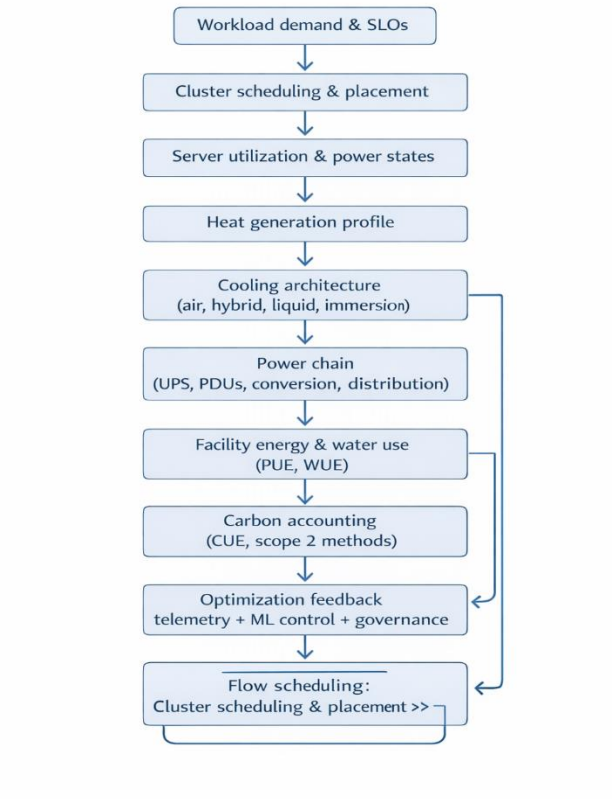
In the EU, the recast Energy Efficiency Directive introduced monitoring and reporting obligations for data centers, supported by a delegated regulation

establishing reporting KPIs and a European database. The Commission's communications specify reporting deadlines and the intent to improve transparency on energy and water footprints (European Commission, 2024).[33] This is a decisive signal: efficiency metrics are no longer only operational tools; they are becoming compliance artifacts—affecting how hyperscale facilities are designed, instrumented, and audited.

V. OPTIMIZATION STRATEGIES

This section presents a practical "stack" of efficiency levers spanning facility and IT. The guiding principle is cascading savings: reducing IT energy (and waste heat) shrinks downstream cooling and power provisioning requirements, amplifying total savings (DOE, 2024).[34]

This diagram reflects the literature's shift toward closed-loop optimization linking scheduling, thermodynamics, and carbon accounting (Masanet et al., 2020; Shehabi et al., 2024; GHG Protocol, 2015).[35]



Cooling technologies and thermal management

Air-side and water-side economization (free cooling). Free cooling reduces compressor runtime by using ambient air or cooling water when outdoor conditions permit, often combined with adiabatic assistance in drier climates. Modelling studies show that free cooling can yield meaningful energy and (sometimes) water tradeoffs across weather regimes, but performance depends on climate, filtration/humidity control, and allowable temperature envelopes (Silva-Llanca et al., 2023; Badiei et al., 2023).[36]

Operating envelopes (ASHRAE) and higher temperature strategies. Raising supply temperatures can reduce chiller lift and increase economizer hours, but must account for server fan response and reliability. ASHRAE's thermal guidance has evolved to define recommended and allowable conditions, and DOE's 2024 guide emphasizes that envelope choices are among the first steps in efficient cooling design (DOE, 2024; ASHRAE, 2021).[37]

Direct liquid cooling (DLC) and hybrid air–liquid. Direct-to-chip cooling uses cold plates and a technology cooling system loop, shifting part of heat removal from air to liquid. This can reduce fan energy and—critically—enable higher coolant temperatures, which improves the feasibility of compressor-less heat rejection (dry coolers, economizers) and increases the quality of recoverable heat (ASHRAE TC 9.9, 2021; Stahlhut et al., 2025; Shehabi et al., 2024).[38]

Immersion cooling. Immersion (single-phase or two-phase) submerges IT equipment in dielectric fluids. Open specifications from the Open Compute Project describe immersion categories and system interface considerations, reflecting the need for interoperable ecosystems at hyperscale scale (Open Compute Project, 2022).[39] Standardization efforts emphasize sustainability requirements for fluids (e.g., low global warming potential, safety constraints), illustrating that thermal efficiency must be co-optimized with environmental and operational risk (Open Compute Project, 2022).[29]

Comparative view.

Cooling approach	Best-fit workload regime	Efficiency mechanisms	Key risks / adoption barriers	Typical metric impacts to watch
Air cooling + containment + economizers	General-purpose compute, moderate densities	Lower fan energy via airflow control; compressor avoidance via economizers	Particulate filtration; humidity control; heat-wave resilience	PUE, CER, fan power, economizer hours
Direct-to-chip liquid (hybrid)	High-density CPU/GPU; AI clusters	Fan reduction; higher coolant temps; better heat capture	Leak management, CDU integration, service procedures	PUE, CER, supply/return temps, pump power
Rear-door heat exchangers	Retrofit-friendly densification	Captures rack exhaust heat; improves air-side effectiveness	Space/weight; facility water coupling	CER, ΔT air, facility water temps
Immersion (single/two-phase)	Extreme densities; modular AI pods	Near-elimination of server fans; high heat transfer; potential heat reuse	Fluid compatibility, maintenance workflows, supply chain	PUE, ERF (heat reuse), reliability/MTTR
Evaporative/adiabatic cooling	Hot/dry climates; cost-optimized designs	High thermodynamic effectiveness with lower compressor use	Water consumption; water-rights constraints	WUE (site), water risk indicators

Sources for table: ASHRAE liquid cooling guidance and TC 9.9 white paper; ISO cooling KPI development (CER); LBNL PUE/WUE modelling discussion of economizers/adiabatic/dry coolers and liquid cooling; OCP immersion specs.[40]

Power chain optimization: UPS, distribution, and conversion losses

As hyperscale power densities rise, electrical losses become a larger absolute quantity even if loss percentages remain similar. Reducing conversion stages and improving part-load efficiency can yield double benefits: less wasted electrical energy and less heat that must be removed (DOE, 2024).[41]

UPS efficiency and operating modes. UPS systems experience efficiency variation with load; strategies include right-sizing, modular architectures, and high-efficiency modes when consistent with reliability requirements. Industry and vendor analyses highlight the importance of efficiency curves and topology choices, but real-world selection must be validated against uptime requirements and failure modes (Active Power, 2023; Shehabi et al., 2024).[42]

Higher-voltage and DC distribution (48 V, 380 V DC, hybrid). The hyperscale ecosystem has increasingly adopted 48 V architectures in racks because higher voltage reduces current, which reduces conduction losses. OCP-related materials note that moving from 12 V to 48 V reduces current draw by a factor of four and conduction losses by a factor of sixteen (Open Compute Project, 2020).[43] Research comparing AC to 380 V DC architectures suggests efficiency advantages for DC under certain configurations and integration scenarios (e.g., PV), though adoption depends on safety standards, equipment availability, and operational practice (Amin et al., 2018).[44]

Renewable integration and grid interaction. Hyperscale operators increasingly match electricity

use with renewable generation through PPAs and other instruments, and leading firms are moving toward hourly carbon-free energy matching. This shift increases the value of (a) flexible load shaping, (b) storage, and (c) geographically distributed workload placement (IEA, 2025; Riepin et al., 2025).[45]

On-site generation (and the reliability–carbon tension). Backup generation is traditionally diesel-based; however, new pathways include hydrogen fuel cells as backup, and in some cases exploration of nuclear or SMR procurement to supply low-carbon firm power through the grid. These strategies are motivated by both reliability and decarbonization, but introduce permitting, technology readiness, and public acceptance constraints (Amazon, 2024; Google, 2024; Microsoft, 2024).[46]

IT-side optimizations: utilization, virtualization, workload placement, and AI scheduling

The most robust finding spanning government guidance and academic synthesis is that utilization-driven consolidation can dominate many facility-side incremental measures, because it attacks energy at the source and triggers cascading savings (DOE, 2024; Masanet et al., 2020).[7]

Virtualization and consolidation. Virtualization increases the number of compute instances per physical server, reducing the number of servers required and therefore reducing IT energy and cooling scale. DOE's 2024 best-practices guide explicitly frames virtualization as a way to "drastically reduce" server count and the size of required cooling equipment (DOE, 2024).[34] Masanet et al. (2020) similarly attributes part of the global decoupling trend to increased virtualization and lower PUE in hyperscale.[23]

Energy-proportional operation and power states. Techniques include dynamic voltage and frequency scaling, aggressive idle states, server power capping, and turning off unused servers—all aligned with the energy-proportional ideal (Barroso & Hölzle, 2007).[47] In hyperscale fleets, practicality hinges on the platform's ability to maintain performance SLOs and resilience under rapid workload changes.

Workload placement and thermal-aware scheduling. Placement decisions affect thermal hotspots and cooling demand. Research in this area increasingly integrates thermal models into scheduling to reduce cooling overhead, especially for dense AI clusters (Shin et al., 2025; Shehabi et al., 2024).[48]

Carbon-aware scheduling and spatio-temporal load shifting. The last five years have produced a surge of work on carbon-aware optimization, including: - temporal shifting of delay-tolerant workloads to periods with lower grid carbon intensity (Wiesner et al., 2021);[49]

- geo-distributed scheduling that incorporates latency and carbon objectives (Souza et al., 2023);[50]
- spatio-temporal load shifting to support 24/7 carbon-free energy matching (Riepin et al., 2025).[51]

Operators have begun translating these ideas into demand-response and load-flexibility programs, which implicitly treat hyperscale data centers as controllable loads supporting grid reliability (Google, 2023).[52]

Facility design: modularity, containment, and airflow management

Airflow management and hot/cold aisle containment. Containment reduces mixing of supply and exhaust air, raises return temperatures, improves cooling unit effectiveness, and can reduce fan power when combined with variable speed drives and airflow control. DOE's best-practices guide describes modular containment systems and the benefits of isolating hot and cold air streams (DOE, 2024).[53]

Modular and prefabricated design. Hyperscale campuses frequently deploy repeatable "blocks" to accelerate construction and improve predictability in cost and schedule. Prefabrication can also enhance commissioning quality and reduce rework, which can indirectly support energy performance by ensuring systems operate closer to design intent (Vertiv, 2019; McKinsey, 2025).[54] Emerging research also explores modular data centers co-located with

renewable generation to reduce transmission needs and enable low-carbon operation, though this remains an evolving approach with siting and grid-integration constraints (Sun et al., 2024).[55]

Monitoring, controls, and maintenance: DCIM, telemetry, and predictive operations

Operational performance gap. A recurring theme in both governmental guides and industry analyses is that design efficiency does not automatically translate into achieved efficiency. Achieved efficiency requires continuous monitoring, commissioning, and control tuning (DOE, 2024).[34]

DCIM and telemetry. Modern hyperscale operations increasingly rely on fine-grained telemetry from servers, network devices, power systems, and cooling plants. This supports real-time optimization, capacity management, and anomaly detection. Even when "DCIM" is defined differently across vendors, the core functionality is converging toward: (a) measurement, (b) modeling, (c) control, and (d) reporting aligned with compliance KPIs (DOE, 2024; European Commission, 2024).[56]

AI-driven predictive control and maintenance. Peer-reviewed work demonstrates AI/ML methods for facility optimization (e.g., reinforcement learning or model predictive control for cooling) and for predictive maintenance in complex building services, suggesting potential to reduce downtime, improve efficiency, and stabilize performance (Lazic et al., 2018; Scaife et al., 2024).[57]

VI. CASE STUDIES AND COMPARATIVE OUTCOMES

This section compiles published, primary operator metrics and notable engineering strategies. Because metrics differ by boundary and averaging conventions, comparisons should be interpreted as indicative, not definitive.

Public KPI disclosures by leading hyperscale operators

Operator	Publicly reported PUE (year, scope)	Publicly reported WUE (year, scope)	Notable efficiency strategies emphasized in primary materials
Google[58]	2024 global fleet average annual PUE **1.09** (global data centers, annual)	Water stewardship disclosures include water-risk sourcing and replenishment, but no single fleet WUE number on the cited pages	Fleet-wide measurement and reporting; emphasis on efficient infrastructure and sustainability operations including water stewardship
Amazon Web Services[59]	2024 global PUE **1.15** (AWS data centers)	No fleet WUE reported in the cited sustainability summary; water-positive commitments exist but measurement disclosure is uneven across the industry	Optimized designs and advanced cooling; claims of 12% more compute and reduced peak cooling energy in new components
Microsoft[60]	Global PUE **1.16 (FY24)** and **1.17 (FY25)** for owned/controlled data centers; regional breakdown published	Global WUE **0.30 L/kWh (FY24)** and **0.27 L/kWh (FY25)**; "zero-water evaporation" design announced for new builds	Higher operating temperatures; free air cooling and rainwater harvesting; exploring hydrogen fuel cells; new "zero-water for cooling" design using closed-loop liquid cooling
Meta Platforms[61]	PUE **1.08** (2022–2024) in environmental data index	WUE **0.19 L/kWh (2024)** in environmental data index	Standardized reporting; emphasis on efficient buildings, renewable energy

Operator	Publicly reported PUE (year, scope)	Publicly reported WUE (year, scope)	Notable efficiency strategies emphasized in primary materials
			purchasing, and water restoration

Sources for table: Google data center efficiency and operating-sustainably pages; Amazon Sustainability Report and AWS sustainability pages; Microsoft datacenter efficiency page and Microsoft Cloud blog; Meta Environmental Data Index.[62]

Interpretation: what case studies imply for the engineering frontier

Three analytically important patterns emerge:

- PUE convergence at the low end. Google's and Meta's published PUE values around ~1.08–1.09 illustrate that best-in-class hyperscale has pushed facility overhead close to practical limits for conventional architectures, consistent with earlier research noting world-class hyperscale near ~1.1 or lower (Masanet et al., 2020).[63]
- Water becomes a binding constraint. Microsoft's explicit reporting of WUE and its move toward "zero-water evaporation" (closed-loop liquid cooling) demonstrates a strategic realignment: evaporative techniques historically helped energy efficiency, but water stress and social license pressure can force design shifts even if they raise PUE slightly (Microsoft, 2024).[64]
- AI densification pushes liquid cooling into mainstream. Government modelling for the U.S. explicitly treats liquid cooling as an emerging technology for dense AI equipment and links higher liquid cooling temperatures to improved efficiency by enabling more "free cooling" and reducing chiller and adiabatic use (Shehabi et al., 2024).[65]

Vendor and ecosystem case evidence: standardization as an efficiency enabler

Hyperscale efficiency is not only about adopting a technology; it is about deploying it safely at scale. Two ecosystem-level examples are instructive:

- Open immersion specifications. The Open Compute Project's immersion fluid specification emphasizes supply-chain and interoperability barriers and positions standardization as enabling adoption and sustainability constraints (e.g., low GWP, safety, long operational lifetime).[29]
- Direct liquid cooling integration challenges. Industry white papers focusing on direct liquid cooling highlight that DLC poses system-level challenges: facility interfaces, CDUs, leak detection, maintenance workflows, and operational maturity—all of which can determine whether theoretical efficiency gains are realized.[66]

Modeling, Quantification, and Carbon Accounting

Energy and water quantification templates

A minimal publishable quantification framework for hyperscale performance should explicitly separate: - IT energy (E_IT): electricity delivered to IT equipment, as defined by the measurement category (ISO/IEC 30134-2).[67]

- Total facility energy (E_total): utility-side energy entering the data center boundary.
- Overhead energy (E_overhead): E_total − E_IT.
- Then: - PUE = E_total / E_IT (ISO/IEC 30134-2).[68]
- WUE = water used / E_IT (ISO/IEC 30134-9; with careful disclosure of site vs source water).[69]
- CUE, analogously, expresses CO_2 intensity of operations per IT energy (ISO/IEC 30134-8).[70]

The LBNL/DOE modelling framework underscores that PUE and WUE vary strongly by cooling system, operating practices, and climate, and that credible scenarios require thermodynamics-based models calibrated to industry-reported ranges (Shehabi et al., 2024).[71]

Carbon accounting: operational versus embodied, and the Scope 2 boundary problem

Scope 2 accounting (market-based vs location-based). The GHG Protocol Scope 2 Guidance

standardizes dual reporting methods: location-based (grid-average emissions where consumption occurs) and market-based (contractual instruments such as RECs and PPAs). This distinction is central for hyperscale because "renewable matching" claims can diverge depending on method (GHG Protocol, 2015).[72]

Operational carbon intensity (CUE) and 24/7 CFE. CUE can be computed with either emissions factor approach; however, the industry trend toward hourly matching requires time-resolved emissions factors and encourages load shifting and storage. Research on spatio-temporal load shifting shows how moving compute across time and location can support 24/7 carbon-free matching but must be balanced against latency, cost, and data sovereignty requirements (Riepin et al., 2025; Souza et al., 2023).[73]

Embodied carbon and LCA. As grids decarbonize and hyperscale operational efficiency improves, embodied emissions from hardware and construction can become a larger share of life-cycle impacts. ISO 14040/14044 provide foundational LCA principles and requirements (ISO, 2006).[74] A recent Nature study applies LCA to advanced cooling technologies at cloud-infrastructure scale, indicating that cooling choices can shift impacts across the stack (server architecture, buildings, and grid interactions), reinforcing the need for life-cycle rather than single-metric optimization (Alissa et al., 2025).[75]

At the operator-methodology level, Microsoft has published an approach (CHEM) for measuring embodied carbon of cloud hardware at fleet scale using process-based LCA combined with granular product data, illustrating how hyperscalers are operationalizing embodied-carbon accounting for procurement and design decisions.[76]

Cost–benefit analysis and decision economics

For publishable cost–benefit comparisons, a common structure is: - CapEx increment for the technology (e.g., CDUs for DLC; immersion tanks; heat-reuse interfaces).

- OpEx savings from reduced energy, water, and maintenance.
- Monetized carbon impacts where applicable (internal carbon price or regulatory).
- Risk adjustments for downtime impact and reliability constraints.

DOE's best-practices framing emphasizes that efficiency measures should be evaluated with both energy and financial impacts (DOE, 2024), consistent with broader critiques that PUE alone is insufficient for investment decisions.[77]

Barriers and constraints

Thermodynamic and physical constraints. As PUE approaches 1.0, remaining overhead is dominated by irreducible physics (fans, pumps, minimal heat rejection) and reliability requirements (redundancy, power conditioning). This is why the frontier moves toward heat reuse, higher-temperature operation, and IT-side reductions rather than only incremental facility tweaks (Masanet et al., 2020; ASHRAE TC 9.9, 2021).[78]

Water constraints and community impacts. Water use creates a social license and regulatory risk; even if evaporative cooling reduces energy, it may not be acceptable in water-stressed regions. Microsoft's design explicitly acknowledges the energy–water tradeoff: avoiding evaporative cooling can increase PUE, but closed-loop liquid cooling reduces water dependence, suggesting water constraints can dominate the objective function (Microsoft, 2024).[64]

Data transparency and comparability. LBNL emphasizes that lack of data availability limits analysis and that greater transparency is needed to support planning, utility engagement, and technology development (Shehabi et al., 2024).[79]
Operational maturity and reliability risk. Liquid and immersion cooling adoption depends on workforce practices (maintenance, leak response), supply chain maturity (fluids, materials compatibility), and standardized interfaces—all highlighted by ASHRAE and OCP materials.[80]

Policy and regulatory considerations

EU reporting and rating scheme. The European Commission has adopted a delegated regulation establishing a first phase of an EU-wide rating scheme and database for data-center energy performance; reporting deadlines were set for September 2024 and then annually (European Commission, 2024).[81] This policy direction implies that hyperscale operators in the EU (and potentially globally, by norm diffusion) must invest in standardized measurement, auditing, and reporting infrastructure and should anticipate potential transitions from reporting to minimum performance standards (European Commission, 2024).[82]

U.S. policy and planning signals. DOE's public statements highlight data centers as a major driver of electricity demand growth and emphasize strategies including flexibility, on-site generation/storage, transmission expansion, and efficient semiconductor technologies (DOE, 2024).[83]

Future research directions

The evidence base suggests six research priorities likely to remain high-value over the next decade:

1. Joint optimization across IT, thermal, and grid layers using real-time telemetry and predictive models, with publishable benchmarks and reproducible datasets (Lazic et al., 2018; Shehabi et al., 2024).[84]
2. Carbon-aware and water-aware scheduling that integrates local water constraints and heat-wave risk alongside carbon intensity and SLO constraints.
3. LCA-integrated design optimization for cooling and power architectures, including embodied-carbon impacts of CDUs, piping, fluids, and modular construction (ISO, 2006; Alissa et al., 2025; Microsoft, 2026).[85]
4. Standardized, auditable metrics for "useful work" (beyond PUE) that are hard to game and meaningful across heterogeneous AI workloads, building on ISO ITEEsv/ITEUsv and related metric research (ISO, 2017; Safari et al., 2025).[86]
5. Heat reuse at scale (district heating, industrial symbiosis) with standardized ERF/energy reuse measurement and net-carbon benefit accounting (ISO, 2021).[87]
6. Grid-interactive hyperscale campuses that can provide demand response, storage, and ancillary services without compromising reliability, especially under AI-induced load variability (IEA, 2025; Takci et al., 2025).[88]

Year	Key Efficiency Inflection Point
2016	ISO/IEC 30134-1 and 30134-2 standardize KPI structure and PUE categories
2016	Early ML-driven cooling optimization publicly reported in hyperscale operations
2020	Global energy-use recalibration highlights decoupling via hyperscale + virtualization
2021	ASHRAE TC 9.9 publishes mainstream liquid cooling guidance
2022	ISO expands KPI set (CUE, WUE) and OCP advances immersion specifications
2024	DOE/LBNL update projects sharp load growth; EU operationalizes data center reporting database
2025	IEA frames AI as key driver; research accelerates on carbon-aware scheduling and load shifting
2026	Embodied-carbon accounting methods scale (e.g., CHEM) and grid-integration research expands

REFERENCES

1. Active Power. (2023). High-efficiency UPS systems for a power hungry world (White paper). [90]
2. Alissa, H., et al. (2025). Using life cycle assessment to drive innovation for sustainable cloud infrastructure. Nature. [75]
3. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2021). 2021 equipment thermal guidelines for data processing environments (reference card). [91]
4. ASHRAE Technical Committee 9.9. (2021). Emergence and expansion of liquid cooling in mainstream data centers (White paper). [27]
5. Badiei, A., et al. (2023). The energy-saving potential of air-side economisers in modular data centres. Sustainability, 15(14), 10777. [92]
6. Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. Computer, 40(12), 33–37. [47]

7. DOE (U.S. Department of Energy). (2024, December 20). DOE releases new report evaluating increase in electricity demand from data centers. [83]
8. DOE (U.S. Department of Energy). (2024). Best practices guide for energy-efficient data center design. [93]
9. European Commission. (2024, March 15). Commission adopts EU-wide scheme for rating sustainability of data centres. [94]
10. European Commission. (2024). Energy performance of data centres. [95]
11. Google. (2024). Power usage effectiveness. [96]
12. Google. (2024). Operating sustainably. [97]
13. Google. (2023, October 3). Using demand response to reduce data center power consumption. [52]
14. Greenhouse Gas Protocol. (2015). Scope 2 guidance. [72]
15. IEA (International Energy Agency). (2025). Energy and AI: Executive summary. [98]
16. IEA (International Energy Agency). (2025). Energy demand from AI. [99]
17. IEA 4E (International Energy Agency – 4E TCP). (2025). Data centre energy use: Critical review of models and results. [24]
18. ISO (International Organization for Standardization). (2006). ISO 14040:2006 Environmental management—Life cycle assessment—Principles and framework. [100]
19. ISO (International Organization for Standardization). (2006). ISO 14044:2006 Environmental management—Life cycle assessment—Requirements and guidelines. [101]
20. ISO/IEC. (2016). ISO/IEC 30134-2:2016 Information technology—Data centres key performance indicators—Part 2: Power usage effectiveness (PUE). [67]
21. ISO/IEC. (2022). ISO/IEC 30134-8:2022 Information technology—Data centres key performance indicators—Part 8: Carbon usage effectiveness (CUE). [70]
22. ISO/IEC. (2022). ISO/IEC 30134-9:2022 Information technology—Data centres key performance indicators—Part 9: Water usage effectiveness (WUE). [102]
23. ISO/IEC. (2021). ISO/IEC 30134-6:2021 Information technology—Data centres key performance indicators—Part 6: Energy reuse factor (ERF). [103]
24. ISO/IEC. (2017). ISO/IEC 30134-4:2017 Information technology—Data centres key performance indicators—Part 4: IT equipment energy efficiency for servers (ITEEsv). [104]
25. ISO/IEC. (2017). ISO/IEC 30134-5:2017 Information technology—Data centres key performance indicators—Part 5: IT equipment utilization for servers (ITEUsv). [105]
26. ISO/IEC. (2023). ISO/IEC 30134-7:2023 Information technology—Data centres key performance indicators—Part 7: Cooling efficiency ratio (CER). [106]
27. Katal, A., et al. (2022). Energy efficiency in cloud computing data centers. Journal of Cloud Computing. [107]
28. Lazic, N., et al. (2018). Data center cooling using model-predictive control. NeurIPS Proceedings. [108]
29. Masanet, E., et al. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984–986. [23]
30. Microsoft. (2024–2025). Measuring energy and water efficiency for Microsoft datacenters. [109]
31. Microsoft. (2024, December 9). Sustainable by design: Next-generation datacenters consume zero water for cooling. [110]
32. Microsoft. (2026). Cloud hardware emissions methodology (CHEM) (White paper). [111]
33. Meta. (2025). Meta environmental data index (includes 2020–2024 PUE/WUE series). [112]
34. Open Compute Project. (2022). Base specification for immersion fluids. [29]
35. Open Compute Project. (2022). OCP immersion requirements rev. 2.0. [113]
36. Open Compute Project. (2020, July 21). Artesyn Embedded Power announces Open Rack version 3 power shelf, evolution to 48-volt infrastructure. [43]
37. Riepin, I., et al. (2025). Spatio-temporal load shifting for truly clean computing. Joule. [51]
38. Safari, A., et al. (2025). A systematic review of energy efficiency metrics for cloud data centers. Electronics. [114]

39. Scaife, A. D., et al. (2024). Improve predictive maintenance through the application of AI: A rapid evidence assessment. Heliyon. [115]
40. Shehabi, A., et al. (2016). United States data center energy usage report (LBNL report). [116]
41. Shehabi, A., et al. (2024). 2024 United States data center energy usage report (LBNL-2001637). [79]
42. Silva-Llanca, L., et al. (2023). Improving energy and water consumption of a data center through air-side free-cooling. Energy and Buildings. [117]
43. Souza, A., et al. (2023). CASPER: Carbon-aware scheduling and provisioning for distributed web services. [50]
44. Stahlhut, M., et al. (2025). Data centers with direct liquid-cooled servers: Higher cooling temperatures, reduced cooling energy, and waste heat reuse. Energy Science & Engineering. [118]
45. Uptime Institute. (2023). Global PUEs—are they going anywhere? [119]
46. Uptime Institute. (2024). Uptime Institute Global Data Center Survey 2024 (report). [120]
47. Wiesner, P., et al. (2021). How temporal workload shifting can reduce carbon emissions in the cloud. arXiv. [49]