Behind-the-Meter Power for Hyperscale Data Centers:
Strategic Advantage or Latency Trap?

George Ralston, Director, Renova Energy Solutions
george.ralston@renovaenergysolutions.com

Technical Insight Paper — 2026

Abstract—*The hyperscale data center industry is reorganising itself around power. Facing multi-year grid interconnection delays, operators are increasingly building dedicated behind-the-meter (BTM) generation — gas turbines, small modular reactors, battery storage, and hybrid microgrids — to bypass congestion and control their own energy supply. The case for BTM is compelling, and growing stronger as the IEA projects global data center electricity consumption will roughly double to 950 TWh by 2030 [1][2]. But the industry's fixation on megawatts is obscuring a critical constraint: the physics of light. The locations that are best for large-scale generation are frequently the worst for low-latency compute. As AI inference becomes embedded in search, industrial control, autonomous systems, and consumer applications, the distance between a data center and its users is no longer a footnote — it is a performance specification. This paper examines the strategic, technical, and operational dimensions of BTM power for hyperscale infrastructure, drawing on published analysis from the IEA, the Open Compute Project (OCP), Schneider Electric, and Vertiv. It argues that power-centric placement is the industry's emerging architectural risk, and that the operators who win the next decade will be those who treat energy, latency, and geographic resilience as co-equal constraints across a layered, distributed infrastructure model.*

Keywords—*behind-the-meter power; hyperscale data centers; AI infrastructure; latency; edge computing; battery energy storage; Open Compute Project; distributed architecture; grid resiliency.*

I. INTRODUCTION

1. The data center industry has always had a power problem. What has changed is its scale. The International Energy Agency (IEA) estimates that global data center electricity consumption reached approximately 415 TWh in 2024 — roughly 1.5% of global electricity use — and projects this will nearly double to around 950 TWh by 2030, with AI identified as the primary driver [1][2]. The numbers are striking. But behind them lies a more fundamental shift: hyperscale operators are no longer simply consumers of electricity. They are becoming energy developers, transmission planners, and infrastructure strategists — and that transformation is forcing a hard rethink of where compute should physically live.

2. Behind-the-meter (BTM) power generation — encompassing gas turbines, small modular nuclear reactors, battery energy storage systems (BESS), fuel cells, dedicated renewables, and hybrid microgrids — has emerged as the industry's primary response to grid constraint. Unable to wait years for utility interconnection, hyperscale operators are building their own power stations. The appeal is obvious: control over capacity, control over timing, and insulation from tariff volatility and transmission congestion.

3. The problem is that power availability and user proximity do not share a map. The locations best suited for large-scale dedicated generation — remote renewable corridors, industrial zones, regions with cheap gas or nuclear access — are frequently the worst locations for low-latency compute. As BTM strategies mature, the industry is discovering that the central question is no longer "How do we power the data center?" but "Where should compute physically exist relative to power, users, and network resilience?" Getting that answer wrong does not just affect performance metrics. For AI inference workloads, it translates directly into degraded user experience, lost revenue, and competitive disadvantage.

4. This paper examines the strategic, technical, and operational dimensions of BTM power for hyperscale infrastructure. It draws on published analysis from the IEA, the Open Compute Project (OCP), Schneider Electric, and Vertiv, and argues that power-centric placement is the industry's emerging architectural blind spot. The winning model of the next decade will not optimise for energy alone. It will treat compute placement, user proximity, network connectivity, and geographic resilience as co-equal design constraints — balanced across a layered, distributed infrastructure model that matches the right workload to the right location.

II. BEHIND-THE-METER POWER: DEFINITION AND SCOPE

5. Traditional data centers connect to the public utility grid through transmission and distribution infrastructure, with utilities providing power capacity, redundancy, and balancing services. Behind-the-meter systems partially or fully bypass this arrangement.

6. In a BTM configuration: the power source is dedicated to the facility; the generation asset may be privately owned or contracted; the facility can operate independently from the broader grid; energy flows directly into the data center campus; and grid interconnection may still exist for backup or

balancing. This turns a hyperscale campus into a private energy ecosystem.

7. Schneider Electric's analysis notes that hyperscale campus plans are now being drawn at up to 11 GW of total capacity at full build-out — over 10% of the Texas grid's record peak — making individual projects national-level energy planning exercises [3]. At that scale, BTM generation is not merely convenient; it is practically necessary.

8. Schneider Electric's research further identifies BESS as a central BTM mechanism, distinguishing between front-of-meter (FTM) utility storage and BTM systems directly accessible to operators. BTM BESS enables peak shaving to reduce costly demand charges, energy independence, and protection against utility supply constraints [4].

III. DRIVERS OF BTM ADOPTION

A. Grid Constraints

9. The largest challenge facing hyperscale expansion is not land — it is power availability. Many regions with strong fiber connectivity and attractive tax environments now face multi-year utility interconnection delays, transmission congestion, limited transformer availability, substation capacity shortages, and rising demand from electrification.

10. Some hyperscale campuses require 500 MW to 1 GW of capacity, with high-density AI compute racks exceeding 100 kW per rack. Google's contributions at the OCP 2025 EMEA Summit outlined a power delivery transformation from 48 VDC to a new +/-400 VDC architecture, enabling IT racks to scale from 100 kW up to 1 MW per rack — a tenfold increase in density that places enormous pressure on upstream supply infrastructure [5]. Behind-the-meter generation allows developers to bypass years of waiting by building generation on-site and controlling commissioning schedules.

B. Energy Cost Predictability

11. Electricity pricing volatility has become a major financial concern. Hyperscale operators increasingly prefer long-term control over fuel sourcing, power purchase structures, energy hedging, and carbon accounting. BTM systems reduce exposure to utility tariff changes, transmission fees, congestion pricing, and capacity market fluctuations. Vertiv's annual data center trends analysis has consistently highlighted power economics and operational stability as defining procurement priorities in the hyperscale and AI markets [6].

C. Improved Resiliency

12. Data centers already rely heavily on redundancy: dual utility feeds, UPS systems, diesel generators, and diverse network paths. BTM power extends this resiliency by reducing dependency on external grid infrastructure. Schneider Electric's research on power supply constraints notes that operators are actively exploring dedicated gas generation and nuclear co-location as a direct response to the inability of utilities to guarantee timely capacity [7].

13. A specific technical challenge has been identified by the IEA: AI workloads create rapid and large swings in power demand, and meeting these reliably can stretch the technical capabilities of on-site gas plants [2]. This underscores the importance of pairing BTM generation with intelligent load management and storage.

D. Faster AI Deployment Cycles

14. AI training workloads often operate continuously at extremely high power densities, unlike fluctuating traditional cloud workloads. Vertiv's deployment of the iGenius project in Italy — delivering a complete AI infrastructure solution combining advanced cooling and power infrastructure for high-density AI — demonstrates the viability of rapidly deploying prefabricated BTM-integrated solutions at scale [8]. The hyperscaler effectively becomes an anchor customer for a private power station.

IV. THE LATENCY CHALLENGE AND THE RISK OF POWER-CENTRIC PLACEMENT

15. The energy conversation dominates industry coverage of hyperscale expansion, and for good reason — power availability is a genuine crisis. But this focus has created a distortion. Compute placement is governed by a matrix of simultaneous constraints: power availability, fiber connectivity, latency to end users, geographic redundancy, regulatory environment, water access, and cooling capability. Power is the most newsworthy of these variables right now. It is not the only one that matters, and in the context of AI inference, it may not even be the most commercially important one.

16. The gravitational pull of BTM infrastructure naturally draws large compute clusters toward locations selected for their energy economics rather than their network position. Rural areas with access to cheap gas or dedicated renewable generation. Industrial zones adjacent to existing power infrastructure. Remote corridors where land is available and permitting is manageable. Each of these choices makes power sense. Each introduces longer network routes, additional optical hops, reduced carrier diversity, and increased round-trip latency to the users and applications that depend on the compute. A data center that wins on power and loses on network position has not solved the problem — it has moved it.

17. The distinction between AI training and AI inference is the crux of this problem, and the industry has not yet fully reckoned with it. Training workloads — the compute-intensive process of building models — are largely asynchronous. They run for days or weeks, tolerate network latency, and can be located wherever

power and cooling are cheapest. Inference workloads — the continuous process of serving those models to users — are the opposite. They are synchronous, latency-sensitive, and user-proximate by necessity. Every millisecond of round-trip network delay between a model and its user is experienced as sluggishness: a search that hesitates, a voice assistant that pauses, a robotics command that arrives late. As AI becomes embedded in industrial control systems, autonomous vehicles, surgical assistance, and real-time financial applications, the latency tolerance of inference workloads approaches zero. A hyperscale campus with excellent BTM generation but poor network position may be perfectly positioned for AI training and structurally unsuited for the inference workloads that will define commercial AI value.

18. Vertiv's data center trends analysis projects a structural shift in workload distribution toward the edge, driven by the latency requirements of smart buildings, distributed energy systems, and 5G applications [9]. This is not a niche trend. As AI inference moves from cloud endpoints to embedded applications — into manufacturing equipment, into hospital systems, into consumer devices — the geographic envelope within which compute can operate becomes progressively tighter. The physics of signal propagation impose hard limits: light travels roughly 200 kilometres per millisecond through fibre, and every router hop adds latency on top. An inference application requiring sub-10ms response time cannot be served from a data centre several hundred kilometres away, regardless of how efficiently that facility is powered. The industry is in the process of discovering that building power infrastructure in the wrong place is not a shortcut — it is a constraint that compounds over time as inference workloads grow.

V. NETWORK DEPENDENCY AND DISTRIBUTED FAILURE RISK

A. Fiber Route Vulnerabilities

19. When compute is centralized around remote BTM generation, the architecture becomes increasingly dependent on long-haul network transport. Even highly redundant fiber systems can fail due to construction damage, cable cuts, subsea faults, fire, flooding, or geopolitical disruption. The farther compute sits from users, the more transport infrastructure becomes mission-critical. In practice, the challenge is often simply transferred from the utility sector to the telecom sector.

B. Carrier Concentration Risk

20. Some remote locations have limited carrier diversity, increasing dependency on specific telecom providers, limited backbone routes, and regional exchange points. A power-rich location may still be network-poor, and this risk compounds with the scale of the facility: the larger the BTM campus, the greater the consequence of any single carrier failure.

C. Cascading Failure Risk

21. If a centralized hyperscale campus fails, traffic must reroute, load balancing becomes more complex, latency spikes occur, and user experience degrades. Distributed architectures naturally absorb failures more effectively. This is particularly important for AI-driven services where uptime expectations approach five-nines availability. The IEA has noted that large, concentrated power loads that scale up rapidly can create special challenges for electricity affordability and grid stability, further motivating distribution rather than concentration [2].

VI. THE ROLE OF OPEN STANDARDS: THE OPEN COMPUTE PROJECT

22. One of the most important enablers of distributed BTM deployment — and one that receives less attention than it deserves — is the standardisation work underway through the Open Compute Project (OCP). Founded in 2011 and now comprising over 400 member companies, OCP's core contribution is making high-density AI infrastructure deployable as a commodity, anywhere in the network [10]. Without standardised power, cooling, and mechanical interfaces, distributed infrastructure is expensive and bespoke. With them, it can be deployed rapidly at any point in the network — from a remote training campus to a metro edge node. OCP's Open Data Center for AI initiative is developing exactly these standardisations, covering power and cooling technologies, management telemetry, and mechanical structure, with explicit focus on enabling high-density AI infrastructure to be deployed flexibly at any tier [11].

23. The technical specifics matter here. OCP's Power Distribution sub-project is defining power densities sufficient to meet the highest rack density requirements, with a roadmap toward low-voltage direct current (LVDC) distribution inside the data hall [12]. An OCP working group including staff from Google, ABB, and Siemens published a 170-page white paper in 2025 making the case for DC power standardisation, arguing that a DC-native approach benefits operators managing both centralised and distributed infrastructure — by reducing conversion losses, simplifying power chains, and enabling more efficient integration of BTM renewable and storage systems [13]. The relevance to distributed BTM is direct: simpler, more efficient power architectures reduce the capital cost and engineering complexity of deploying AI infrastructure at the edge, where bespoke design is prohibitively expensive.

24. The OCP's alliance with Current/OS, formed in December 2025, reinforces this direction [14]. By enabling direct coupling of DC power sources to the DC busbar with zero or one conversion step, and by

providing a migration path from today's 48V busbar racks to +/-400V and 800V architectures, the work makes distributed BTM deployment more economically tractable. The implication for the latency argument is significant: as standardisation reduces the cost premium of edge-capable BTM infrastructure, the barrier to deploying inference compute closer to users — with locally sourced power — falls substantially.

VII. VENDOR PERSPECTIVES: SCHNEIDER ELECTRIC AND VERTIV

A. Schneider Electric

25. Schneider Electric, through its EcoStruxure platform and Secure Power division, has developed one of the most comprehensive frameworks for layered data center infrastructure. Its analysis describes a hybrid data center architecture of centralized, regional, and local edge tiers, with distinct power and connectivity requirements at each layer [15]. This tiered model aligns closely with the BTM distributed architecture thesis: large BTM campuses for training and bulk compute; regional hubs for redundancy; and local EcoStruxure-managed edge deployments for inference and real-time analytics.

26. Schneider Electric's CTO, Jim Simonelli, argued at Data Center World 2026 that the industry's debate over AC versus DC misses the bigger constraint: the rack's internal real estate. Higher-voltage DC is emerging less as a pure efficiency play and more as a way to reclaim rack space for GPUs — a constraint that equally applies to BTM-powered remote campuses and distributed edge nodes [16].

B. Vertiv

27. Vertiv, as a global provider of critical digital infrastructure across the power and cooling spectrum, has positioned its product strategy explicitly around the transition from centralized cloud to distributed AI-edge infrastructure. Its portfolio of power, cooling, and IT infrastructure solutions extends from the cloud to the edge of the network, and its 2025 acquisition of Great Lakes Data Racks & Cabinets was specifically aimed at strengthening pre-engineered, AI-ready rack solutions for enterprise, edge, colocation, and hyperscale markets [17].

28. Vertiv's prefabricated modular data centers (PMDCs), developed in partnership with Dell Technologies, represent a practical implementation of the BTM-distributed vision: pre-engineered power and cooling infrastructure that can be deployed rapidly at any point in the network — from a remote BTM generation campus to a metro edge inference node — without requiring bespoke facility design [18].

VIII. THE CASE FOR DISTRIBUTED BEHIND-THE-METER ARCHITECTURE

29. The future is unlikely to be a binary choice between grid-connected centralized hyperscale and remote BTM mega-campuses. The winning model will combine regional hyperscale campuses, distributed edge infrastructure, localized BTM systems, flexible microgrids, and intelligent workload distribution. The industry is effectively rediscovering a principle long understood in critical infrastructure engineering: highly centralized systems optimize efficiency, while distributed systems optimize survivability.

30. Geographic concentration creates systemic risk. A single massive BTM site may contain hundreds of megawatts of compute, large AI clusters, vast data storage volumes, and critical cloud infrastructure. Natural disasters, grid instability, cooling failures, or political disruption can simultaneously impact enormous amounts of digital infrastructure. Distributed architectures reduce this exposure by allowing users to dynamically connect to nearby compute nodes, alternative regional hubs, or secondary inference clusters.

31. The sustainability argument for BTM also has a distributed dimension. The IEA notes that renewables remain the fastest-growing source of electricity for data centers, growing at 22% annually between 2024 and 2030, yet natural gas and coal together are still expected to meet over 40% of additional electricity demand from data centers until 2030 [19]. Distributed BTM systems, particularly when co-located with local renewable generation and BESS, can improve renewable integration and reduce transmission losses in ways that remote mega-campuses cannot.

IX. A LAYERED INFRASTRUCTURE MODEL

32. The most probable long-term outcome is a layered infrastructure model with three distinct functional tiers, each with distinct BTM power requirements and network connectivity profiles.

A. Layer 1 — Centralised AI and Cloud Campuses

33. These facilities consume enormous power levels, use large-scale BTM generation, focus on AI training and bulk processing, and prioritise energy efficiency and cost. They may be located where power is abundant, land is available, cooling is feasible, and fiber backbones already exist. BTM generation at this tier is most clearly justified: the power economics are decisive, and the workloads are tolerant of geographic separation.

B. Layer 2 — Regional Data Hubs

34. Regional facilities provide redundancy, reduce interregional latency, support cloud resiliency, enable disaster recovery, and balance traffic loads. They become critical network stabilisation points. Both Schneider Electric and Vertiv identify this tier as essential for managing the reliability and performance expectations of modern AI-driven services.

C. Layer 3 — Edge and Metro Compute

35. Edge infrastructure supports AI inference, real-time analytics, smart manufacturing, connected mobility, autonomous systems, and consumer AI applications. These facilities do not require gigawatt-scale generation; they benefit from strategic placement, local resiliency, low-latency network access, and distributed microgrid capability. Smaller-scale BTM solutions — from Vertiv's prefabricated modular systems to Schneider Electric's EcoStruxure-managed edge deployments — are particularly well-suited to this tier.

X. THE CONTINUING ROLE OF UTILITIES

36. Some industry narratives imply that hyperscalers can completely bypass utilities. That is unlikely. Even advanced BTM deployments still depend heavily on grid synchronisation, backup imports, frequency stability, market balancing, transmission support, and ancillary services. Utilities will remain essential partners. The difference is that hyperscalers increasingly want greater control over capacity timing, generation mix, resiliency architecture, and operational flexibility — shifting the relationship from passive customer to active infrastructure participant.

37. As Utility Dive has noted in its analysis of the emerging 'bring your own power' strategy among data center operators, the future requires a partnership model where the grid provides connectivity and market co-ordination, while large customers bring distributed resiliency and flexibility — because the grid cannot manage this transition alone [20].

XI. CONCLUSION

38. Behind-the-meter power is not a passing trend. It is a structural response to a structural problem: the utility grid cannot keep pace with the speed at which hyperscale AI infrastructure needs to expand. Grid congestion, multi-year interconnection queues, limited transformer availability, and rising electrification demand have made BTM generation a practical necessity for operators at the frontier of AI compute. The IEA's projection that global data center electricity consumption will roughly double by 2030 only strengthens the case [1][2].

39. The risk is not that operators will build BTM infrastructure. They will, and they should. The risk is that the industry will optimise too heavily around power availability and treat latency as a secondary concern — solvable later, addressable through network engineering, someone else's problem. It is not. AI inference is the primary mode through which AI creates economic value for end users, and inference is fundamentally proximity-constrained. The further compute sits from users, the more every millisecond of latency degrades the experience. An industry that builds its next generation of infrastructure around power economics alone is not solving the compute placement problem — it is creating a new version of it.

40. The architecture that resolves this tension is not a binary choice between centralised BTM mega-campuses and fragmented edge deployments. It is a layered model: large energy-optimised facilities for training and bulk compute, where geographic separation is acceptable and power economics are decisive; regional hubs for redundancy and interregional traffic management; and distributed edge infrastructure for inference, real-time analytics, and the latency-sensitive applications that will define AI's commercial footprint. Open infrastructure standards from OCP, modular deployment capabilities from vendors such as Schneider Electric and Vertiv, and a partnership model with utilities that positions hyperscalers as active infrastructure participants rather than passive customers — these are the building blocks of that architecture.

41. The operators who define the next era of digital infrastructure will be those who resist the temptation to optimise around a single constraint. Power, proximity, and resilience are not a hierarchy — they are a system. The facilities that get built in the right places, powered in the right ways, and connected with the right network architecture will not just outperform their competitors on efficiency metrics. They will be the ones still standing when the workloads that matter most — low-latency inference at scale — arrive in earnest. Megawatts matter. So do milliseconds.

XII. REFERENCES

[1] International Energy Agency (IEA). Energy and AI. April 2025. Available: https://www.iea.org/reports/energy-and-ai

[2] International Energy Agency (IEA). Key Questions on Energy and AI. April 2026. Available: https://www.iea.org/reports/key-questions-on-energy-and-ai

[3] Schneider Electric Blog. "How hyperscaler growth is rewriting the data center playbook." December 2025. Available: https://blog.se.com/datacenter/2025/12/08/how-hyperscaler-growth-is-rewriting-the-data-center-playbook/

[4] Schneider Electric Blog. "The future of data centers: Battery Energy Storage Systems (BESS)." May 2024. Available: https://blog.se.com/datacenter/2024/05/01/the-rise-of-bess-powering-the-future-of-data-centers/

[5] Google Cloud Blog. "Enabling 1 MW IT racks and liquid cooling at OCP EMEA Summit." April 2025. Available: https://cloud.google.com/blog/topics/systems/enabling-1-mw-it-racks-and-liquid-cooling-at-ocp-emea-summit

[6] Vertiv Holdings Co. Data Center Trends 2025. December 2024. Available: https://www.vertiv.com

[7] Schneider Electric Blog. "Strategies for data centers facing power supply constraints." December 2025. Available: https://blog.se.com/datacenter/2025/12/03/solving-power-puzzle-strategies-data-centers-supply-constraints/

[8] Vertiv Holdings Co. Q1 2025 Earnings Release. April 2025. Available: https://investors.vertiv.com

[9] Vertiv Holdings Co. "2022 Data Center Trends to Watch." December 2021. Available: https://www.vertiv.com

[10] Open Compute Project Foundation. Wikipedia overview. Available: https://en.wikipedia.org/wiki/Open_Compute_Project

[11] Open Compute Project Foundation. "Realizing the Open Data Center Ecosystem Vision." October 2025. Available: https://www.opencompute.org/blog/realizing-the-open-data-center-ecosystem-vision

[12] Open Compute Project Foundation. Power Distribution Project. Available: https://www.opencompute.org/projects/power-distribution

[13] Data Center Dynamics. "OCP members tout DC power in the data center to meet growing AI energy demands." May 2026. Available: https://www.datacenterdynamics.com/en/news/ocp-members-tout-dc-power-in-the-data-center-to-meet-growing-ai-power-demands/

[14] Open Compute Project Foundation. "OCP and Current/OS Form New Alliance." December 2025. Available: https://www.opencompute.org/blog/open-compute-project-foundation-and-currentos-form-new-alliance

[15] Schneider Electric. "Data Center Sustainability and Planning." Available: https://www.se.com/us/en/work/campaign/data-centers-of-the-future/

[16] Data Center Knowledge. "Data Center World 2026: Power Architecture Pushed Beyond the Rack." April 2026. Available: https://www.datacenterknowledge.com/build-design/data-center-world-2026-new-limits-push-power-architecture-beyond-the-rack

[17] Vertiv Holdings Co. "Vertiv Completes Acquisition of Great Lakes Data Racks & Cabinets." August 2025. Available: https://www.vertiv.com

[18] Vertiv / Dell Technologies. "Modernizing federal IT with AI-ready modular data centers." White Paper, 2025. Available: https://www.vertiv.com/globalassets/documents/white-papers/vertiv-dell-federal-ai-and-it-modernization-wp-en-na-sl-80228-web.pdf

[19] International Energy Agency (IEA). "Energy Supply for AI — Energy and AI." April 2025. Available: https://www.iea.org/reports/energy-and-ai/energy-supply-for-ai

[20] Utility Dive. "Why data centers will need a 'bring your own power' strategy." March 2026. Available: https://www.utilitydive.com/news/why-data-centers-will-need-a-byop-strategy-bring-your-own-power/812004/